AD________________

Award Number: W81XWH-05-1-0031

TITLE: High-resolution Mapping of Structural Mutations in Prostate Cancer with Single Nucleotide Polymorphism Arrays

PRINCIPAL INVESTIGATOR: Dr. Rameen Beroukhim

CONTRACTING ORGANIZATION: Dana-Farber Cancer Institute
Boston. MA 02115

REPORT DATE: November 2006

TYPE OF REPORT: Annual Summary

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for Public Release;
Distribution Unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE
01-11-2006

2. REPORT TYPE
Annual Summary

3. DATES COVERED
1 Nov 2004 – 31 Oct 2006

4. TITLE AND SUBTITLE

High-resolution Mapping of Structural Mutations in Prostate Cancer with Single Nucleotide Polymorphism Arrays

5a. CONTRACT NUMBER

5b. GRANT NUMBER
W81XWH-05-1-0031

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Dr. Rameen Beroukhim

E-Mail: rameen_beroukhim@dfci.harvard.edu

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Dana-Farber Cancer Institute
Boston. MA 02115

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES
Original contains colored plates: ALL DTIC reproductions will be in black and white.

14. ABSTRACT

NOT PROVIDED

15. SUBJECT TERMS Prostate cancer, genomics, chromosome structure, cancer progression and metastasis, single nucleotide polymorphisms, genotyping digital karyotyping cytogenetics, loss of heterozygosity, oligonucleotide, array

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** USAMRMC
a. REPORT U	**b. ABSTRACT** U	**c. THIS PAGE** U	UU	55	**19b. TELEPHONE NUMBER** *(include area code)*

Table of Contents

A. Introduction

The proposal for DOD Award #W81XWH-05-1-0031 focused on the systematic mapping of large-scale genetic alterations in prostate cancer, and relating these mutations to prostate cancer progression. To that end, the proposal suggested the application of single nucleotide polymorphism (SNP) array technology to characterize large-scale genetic alterations in the prostate cancer genome.

1. High-resolution single nucleotide polymorphism arrays

SNPs are the most common genetic variation in the human genome; more than 6,000,000 have been identified (2). The use of single-nucleotide polymorphisms to study the germline genetic susceptibility to disease is well appreciated and an evolving technology designed to conduct such studies is the use of oligonucleotide arrays to interrogate these SNP markers in a high-throughput, highly parallel fashion (3-5). These oligonucleotide arrays specifically detect the two different alleles of each SNP (Figure 1). The most advanced commercially available 100K arrays detect 116,204 SNPs. With a median intermarker distance of 8.9 kb, this represents greater than 5 SNPs per gene, affording state-of-the art resolution for large-scale genotyping purposes (6).

To prepare target for the 100K arrays, genomic DNA is digested with XbaI or HindIII (in separate reactions). HindIII and XbaI linkers are ligated and single-primer PCR amplification is carried out to amplify fragments ranging from 200-2000 bp, resulting in a partial genome representation. The fragments are labeled with streptavidin, fragmented and hybridized to arrays that contain the 58,000 probe sets for either the XbaI or HindIII digest. The probe set for each SNP consists of 10 perfect match (pm) probes to each allele, along with 10 mismatch (mm) probes, for a total of 40 probes. A detailed description of the protocols and technology for these 100K SNP arrays is available at www.affymetrix.com/support/technical/datasheets/100k_datasheet.pdf.

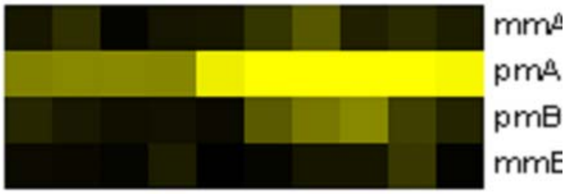


Figure 1: View of a probe set for a single SNP showing a homozygous "A" call. For each SNP, 40 oligonucleotide probes are tiled onto SNP arrays interrogating 116,204 SNPs across the genome. These include perfect match (pm) and mismatch (mm) probes directed against each allele. The SNP position slides 5' to 3' among the pm or mm probes to each allele. The fluorescence pattern indicates which alleles are present; the fluorescence intensity indicates the quantity of bound DNA.

The scale and precision with which high-density SNP arrays interrogate independent alleles prompted our group (led by William Sellers and Matthew Meyerson) to spearhead efforts applying this technology to the analysis of somatic genetic alterations present in human malignancies. Several features of SNP arrays suggested that they might constitute an ideal platform for cancer genomic analyses: 1, determination of allele status across cancer genomes provided a basis for large-scale, high-precision loss of heterozygosity (LOH) analysis; 2, probe set hybridization yielded a signal whose intensity also reflected the copy number at that locus; and 3, the resolution afforded by SNP array marker densities exceeded that of most CGH options.

a. High-resolution loss of heterozygosity analysis

The somatic conversion of heterozygous germline alleles to a homozygous state (LOH) may occur through hemizygous deletion alone (resulting in concomitant copy loss) or followed by gene duplication (copy-neutral LOH). Interestingly, copy-neutral LOH, which is undetectable by conventional CGH methods, represents up to 80% of LOH events in some tumor sets (7), and the primary mechanism of LOH in particular genomic regions of individual cancer types (8, 9). Considerable experimental evidence supports the notion that LOH represents a key mechanism for tumor suppressor inactivation. Indeed, nearly all common tumor suppressor genes occur in regions that frequently undergo LOH (prominent examples include p16, PTEN, pRB, and p53).

Published data by the Sellers and Meyerson groups and by others demonstrate that SNP arrays provide high-resolution maps of LOH when one compares the pattern of heterozygosity in the constitutional germline DNA to the pattern seen in the tumor (10-20). More recently, we have developed methods of analyzing homozygous allele frequencies and regions of linkage disequilibrium to map regions of LOH without the use of paired normal germline DNA samples (21). This has allowed us to map LOH in cell lines and xenografts and to determine the similarity or differences in this data compared to authentic human tumors.

b. Genome-wide maps of copy number aberrations

Our group and others have found that comparison of signal intensities derived from each SNP probe (instead of allele call data) to corresponding signal data from normal genomes allows determination of copy number changes present within tumor samples (22, 23). The concordance with quantitative PCR has generally been excellent, though high-level copy number gains are often underestimated on SNP arrays, presumably due to saturation effects. Various cancer genomes are now beginning to be mapped in this way (24, 25), including analyses by our group, using 100K arrays, of the lung cancer genome (26) and of the NCI60 cell line set (27). The high resolution of the 100K arrays allowed the discovery, in this latter case, of the novel oncogene MITF in melanoma cell lines and metastatic samples.

To that end, the specific aims proposed were:

1. **To isolate DNA from 50 localized and 50 metastatic prostate cancers after laser-capture microdissection, along with DNA from corresponding germline tissue.**
2. **To generate genome-wide high-resolution maps of LOH and copy-number alterations using SNP arrays containing probes for 100,000 markers.**
3. **To identify and validate candidate somatic genetic alterations differing in prevalence between localized and metastatic cancers, and develop markers for clinical association studies.**

Significant progress has been made in all 3 specific aims. However, due to unanticipated difficulties with respect to Specific Aim 1, the number of tumors analyzed to date is smaller than the number set out in Specific Aim 1, and completion of the validation studies in Specific Aim 3 is still underway. The progress and difficulties will be outlined in the next section.

B. Body

1. Specific aim #1: To isolate DNA from 50 localized and 50 metastatic prostate cancers after laser-capture microdissection, along with DNA from corresponding germline tissue.

Reconstitution experiments have shown (11) that contamination of cancer cells with greater than 10% normal cell genomes results in a significant degradation in the ability to determine LOH. Prostate cancers tend to have large concentrations of intervening stroma. Thus, to apply SNP array technology to the study of prostate cancer, samples must be enriched for tumor. In this aim, we attempt to preserve the detection of both LOH events and copy number changes in prostate samples using laser capture microdissection (LCM)-based methods for tumor enrichment.

Our experience has shown that LCM of 2 mm^2 of prostate tissue takes between 2-4 hours and yields 50-100 ng of DNA. A 100K SNP array set requires 500 ng of DNA; to produce this amount of DNA on a large number of tumors quickly becomes prohibitively time-consuming. Fortunately, several methods of whole genome amplification (WGA) exist (28). Among the most promising of these is multiple displacement amplification (MDA) (29), which makes use of a polymerase with exonuclease activity and random primers to perform isothermal amplification, with yields as high as 10,000-fold or greater. As opposed to PCR-based methods, the DNA produced has long fragment lengths and low error rates. We have shown (16) that, using 10 ng of high-quality template DNA, one can produce tens of micrograms of DNA with MDA methods using the Φ29 polymerase. The DNA product preserves genotyping information with 99.8% accuracy, and copy numbers determined from this DNA are 87% concordant with the unamplified template DNA. Much of the 13% discordance in copy number estimates was not functionally important, as it was due to lower saturation levels in WGA DNA—meaning very high amplifications (copy number 6 or greater) were not seen to be as high in WGA as unamplified DNA—although they were noted to be high in both groups. Therefore, we have attempted to apply MDA to DNA obtained from laser-capture mcirodissected tissue

This section will describe progress in 3 sub-aims:

a. Characterization of LCM and WGA conditions for optimal reproducibility
b. Data generation from primary and metastatic tumors

a. Characterization of LCM and WGA conditions for optimal reproducibility

Early on, we found that DNA from LCM can also serve as a template for WGA, providing product that gives similar results on SNP arrays to unamplified DNA. Among four highly enriched tumors, high-density SNP array data was obtained after either macrodissection (either a 2 mm cubic biopsy of tissue, or tissue needle-dissected from a glass slide) or laser capture microdissection. Overall call rates, reflecting the percentage of SNPs for which genotypes could be assigned, averaged 93.7% and 93.6% for macrodissected and microdissected tissue, respectively. Moreover, concordance rates between genotype calls from macrodissected and microdissected tissue averaged 98.2%. Although this concordance rate is slightly worse than that obtained with the highest quality DNA (99.85% in our hands (16)), it is high enough to accurately assign regions of LOH (Figure 2).

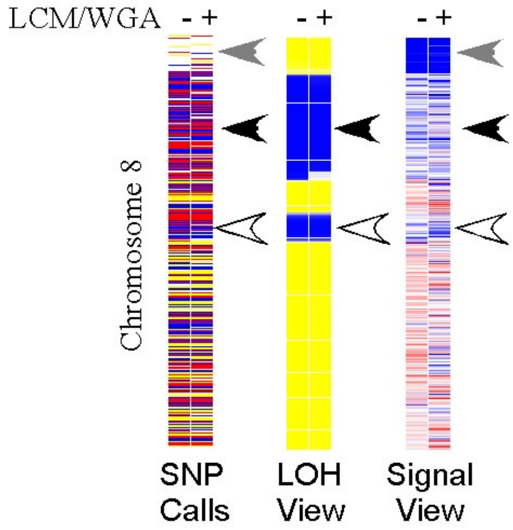


Figure 2: High-quality SNP array data generated after laser capture microdissection and whole genome amplification. Top: Data generated from a highly pure tumor after macrodissection is displayed alongside data from the same sample after LCM and WGA. Genotype calls are displayed in the left panel (yellow – AB (het), red – AA, blue – BB, white – No Call), LOH inferences in the middle panel (blue – LOH, yellow – retention), and signal intensities in the right panel (red – high, white – normal, blue – low). A homozygous deletion on chr. 8 (gray arrow) and regions of LOH (black and white arrows) are indicated.

Likewise, the ability to identify copy number aberrations is preserved (Figures 2,3). The main concern here is due to the potential uneven nature of amplification by WGA. In fact, we know that certain regions of the genome are better represented in WGA product than others, with up to 6-fold variations between different regions (29). As long as these biases are consistent, however, normalization against control samples that have undergone whole genome amplification under similar conditions will correct for them, leaving only the underlying signal intensity changes reflecting copy number aberrations inherent in the sample. To test how consistent these biases are along a range of WGA conditions, DNA obtained from laser capture microdissected benign prostate tissue was amplified under varying conditions.

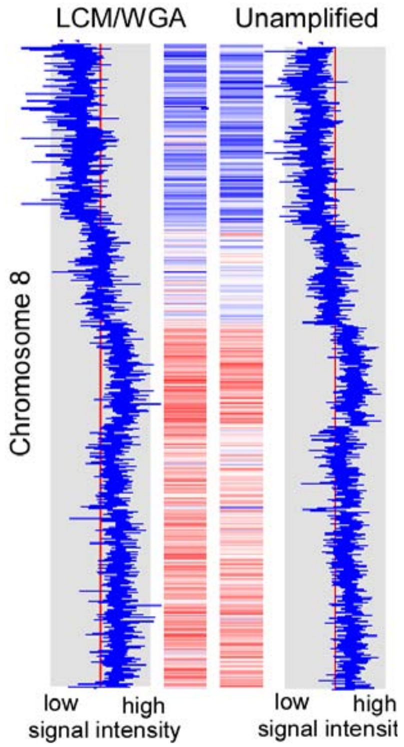


Figure 3: Preservation of signal intensity variations after laser capture microdissection and whole genome amplification. Normalized signal intensities for a second highly pure tumor are displayed (as in Figure 5, right panel), along with graphs of those signal intensities on either side. Both amplifications and deletions appear highly reproducible after laser capture microdissection and whole genome amplification.

Specifically, as the amount of WGA template was increased from 4 ng to 64, the SNP array signal intensities became more similar to unamplfiied controls (Table 1). However, signal intensities from LCM/WGA were highly consistent, when compared against DNA undergoing LCM/WGA under similar conditions (mean variance 0.11, same as unamplified controls; Table 1). Moreover, this consistency remained even when DNA from LCM was whole genome amplified on different days (variance = 0.12), or using different lots of polymerase, dNTP, and random primers (variance = 0.10). Finally, the signal intensities from LCM/WGA appear robust to 2-fold changes in amount of DNA template (variance = 0.13).

With optimal conditions for LCM and WGA determined in benign prostate tissue, we initiated data collection by performing LCM on primary tumors along with germline tissue from paired seminal vesicles.

Table 1. Mean variance between normalized SNP array signal intensities obtained from WGA product using various amounts of template DNA, versus signal intensities from unamplified DNA. Template DNA used for WGA was produced from LCM of benign prostate tissue.

Template amount	4 ng	8 ng	16 ng	32 ng	64 ng	Unamplified*
Variance from unamplified DNA	0.30	0.26	0.24	0.23	0.23	0.11
Variance from similarly prepared DNA	-		-		-	0.11

*Variance between normalized signal intensities from repeat SNP arrays using the same unamplified DNA

Unfortunately, we soon came to find that histology affects the quality of DNA obtained after LCM and WGA. For instance, in one experiment we performed LCM and WGA on 3 primary prostate cancers along with paired uninvolved seminal vesicles. In each case, 32 ng of DNA was used from the laser captured cells as template for WGA, and 250 ng of WGA product was used for restriction digest, amplification, and hybridization to 50K Xba arrays. WGA was performed using the same reagents and at the same time, for all samples. However, genotyping call rates were excellent for the germline DNA obtained from the seminal vesicles, ranging from 95-98%, and were low for the tumors, ranging from 85-90%. Moreover, signal intensity profiles were much noisier for the tumor DNA (data not shown), precluding high-resolution copy number analysis.

As both the tumor and normal seminal vesicle were resected from the patient simultaneously, the cause of the difference in DNA quality between them is likely due to either 1) the differing tissue characteristics between the tumor and normal tissue, or 2) differences in the laser capture process itself. In the case of 2), we recognized that small nests of cancer cells have to be captured from the tumor, whereas large regions of normal tissue could be captured. The result is that the captured tumor cells, on average, lie closer to the line cut by the laser. Therefore, we hypothesized that the laser was causing direct damage to the DNA. In fact, when we obtained DNA from benign prostate tissue after conducting LCM using small shapes such that all regions collected lay close to the laser cut line, we found that SNP call rates decreased (Table 2).

To improve on these results, we purchased an Arcturus Veritas laser capture microdissection machine, allowing LCM to be conducted by capturing cells individually using an IR laser whose energy is focused on melting a polymer that sticks to the cells, thereby reducing the amount of energy going into the tissue itself. As expected, use of the infrared laser improved the call rate for DNA obtained from small shapes, although it appeared to lead to worse results when large shapes are used—possibly due to overall heating of the tissue when the IR laser is applied to large regions (Table 2).

Table 2. SNP array call rates as a function of LCM shape size and laser used.

Shape size (diameter)	Laser	Call rate
0.5 um	UV	81%
0.5 um	IR	92%
1.5 um	UV	92%
1.5 um	IR	83%

Using the infrared laser for LCM of small nests of tumor, we then tested whether use of the infrared laser could enable us robustly to obtain high-quality SNP array data from primary prostate cancer tissue. Unfortunately, despite robust success in benign tissue, SNP array call rates obtained from primary prostate cancers frequently ranged as low as 81%.

b. Data generation from primary and metastatic tumors

Through the Gelb Center at the Dana-Farber Cancer Institute, we have IRB- approved access to several hundred fresh frozen primary prostate cancers, along with uninvolved seminal vesicles. Metastatic tumors have been obtained from several sources, including hormone-naïve lymph node metastases from Dr Mark Rubin, further metastatic tissue through the rapid autopsy programs at the University of Michigan (aided by Drs Rubin and Kenneth Pienta) and University of Washington (aided by Dr Lawrence True), and bone metastases through Drs Steven Balk and Dr Mary-Ellen Taplin. Due to the difficulties with LCM followed by WGA cited above, we aimed in LCM to obtain sufficient unamplified DNA for SNP array analysis. We found this approach, although time-consuming, reliably provides high-quality data. As of this time, we have high-quality, high-resolution SNP array data on 84 prostate tumors, including 24 primary tumors, 38 metastatic tumors, and 22 model systems.

2. Specific aim #2: To generate genome-wide high-resolution maps of LOH and copy-number alterations using SNP arrays containing probes for 100,000 markers.

We currently have a comprehensive analysis of the data generated from 63 of these tumors.

a. Generation of LOH maps

In the application for this award, we described a method we had developed to identify regions of LOH without the use of paired normal DNA. Although we are obtaining paired normal DNA for all primary and metastatic tumors in this study, we also have SNP array data from prostate cancer model systems for which paired normal DNA is unavailable. The method used to determine LOH without paired normal DNA was originally developed using data from SNP arrays probing 10,000 loci throughout the genome. When applying the method to 100K SNP array data, we found that the haplotype structure of the genome reduced the specificity of the method, and improved the method to take this haplotype structure into account. This method has now been published (21); Appendix I.

b. Generation of copy-number maps

For the generation of copy-number estimates, both systematic and random errors in signal intensity data have to be minimized. We found that the main source of systematic error is *batch effect*, whereby a batch of samples that simultaneously undergo DNA digestion, amplification, labeling, and hybridization to arrays, will have similar signal intensity alterations (high or low signal intensity) compared to samples processed at other times. This, in turn, leads to the appearance of amplicons and deletions restricted to that batch (Figure 4). Strict control of experimental conditions and normalization against reference samples from the same batch can minimize, but tend not to eliminate, these batch effects. In turn, these batch effects can lead to the identification of spurious regions of recurrent amplification and deletion.

We posited that the identifying characteristic of an alteration due to batch effect is that the alteration consistently occurs within one batch, and consistently does not occur within other batches of similar samples. Therefore, for each batch containing at least 5 samples, we identified the distribution of signal intensities for each SNP, and compared this using a T-test with the distribution of signal intensities in all other batches. When the p value was less than 0.001, we considered that the SNP had undergone a systematic alteration due to batch effect, and subtracted a constant amount from the signal intensities at that SNP for all samples in the batch, so that the mean signal in that batch equaled the mean signal in all other batches (Figure 4).

Whereas systematic errors can lead to the identification of spurious regions of amplification and deletion, random errors tend to reduce our sensitivity to identifying regions of real importance. Most importantly, the error in the signal intensity measured at a given SNP can lead to that SNP being spuriously identified as amplified or deleted, leading to downstream errors in estimates of the frequency of lesions at that SNP locus. A variety of smoothers, developed for CGH data, reduce noise levels at each locus by involving information from neighboring loci (30). We have found GLAD (31), which identifies segments with a constant copy number and averages the signal intensities across all loci in each segment, provides the most accurate results in a reasonable amount of computational time (data not shown). Several alternative software packages (dChipSNP, CNAT, CNAG, GIM) (22, 23, 32, 33) also exist to convert probe-level data into overall SNP-specific signal intensities. Preliminary results seem to point to CNAG as producing the most optimal signal-to-noise ratios (data not shown).

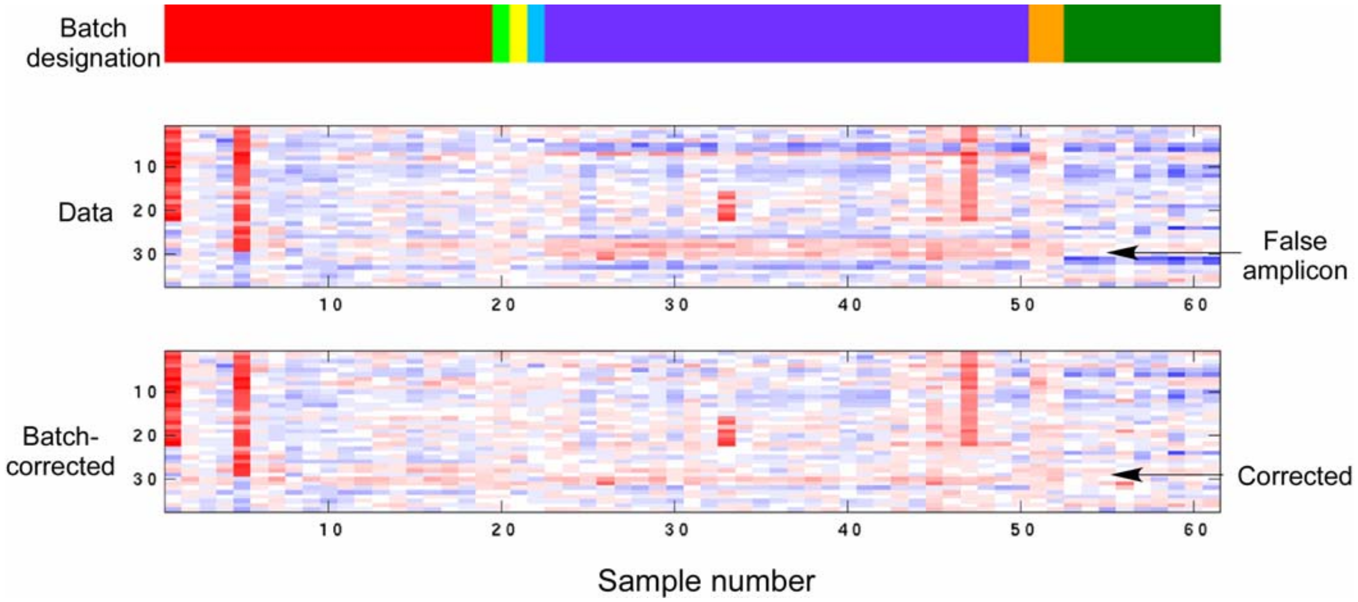


Figure 4: Batch effect and correction. Signal intensity data are displayed for 61 samples, each as a column. These samples were run in 7 batches, designated by different colors in the top panel. The normalized intensities for a set of 40 consecutive SNPs are displayed (as in Figure 2) in the middle panel. In the batch designated in blue (top panel), several adjacent SNPs appear to have consistently high signal intensities, giving the false appearance of a recurrent amplicon. Data corrected for this batch effect are displayed in the bottom panel.

c. Identification of significant chromosomal events

We posited that information as to the importance of a region in sustaining cancer lay not only in the frequency with which that region undergoes lesions, but also the also in the amplitude of the lesions that occur. Therefore, we designed scores for amplification and deletion that included both sources of information. Namely, for each SNP locus we calculate:

$$\text{Amp} = f_{amp} \times \log_2(\hat{S}_{amp}), \text{ and}$$
$$\text{Del} = f_{del} \times -\log_2(\hat{S}_{del}) \qquad (1)$$

where f_{amp} and f_{del} represent the frequency of amplification and deletion, respectively, and $\hat{S}_{amp}$ and $\hat{S}_{del}$ represent the average normalized signal intensity of samples with amplifications and deletions. We scored LOH by the frequency alone.

The significance of each particular Amp, Del, or LOH score is then determined by comparing it to similar scores determined from all permutations of the data, allowing the calculation of p values and, to correct for multiple hypothesis testing, False Discovery Rate (FDR) q values (34).

When applied to our data from 63 prostate cancers, we obtained the Amp, Del, and LOH scores displayed in Figure 5. Regions of amplification and deletion having q values less than 0.01 (i.e. having less than a 1% probability of occurring by chance alone) were designated as significant. The region surrounding the androgen receptor is the most significantly amplified, whereas the region surrounding PTEN is the most significantly deleted.

Multiple other regions of significant amplification and deletion are seen. Within these, the most likely locations of the targeted oncogenes or TSGs were felt to be the regions of minimal q value. Table 3 shows the boundaries of some of these and compares them to locations of known oncogenes and TSGs. The accuracy in identifying these known targets is remarkable, and suggests the candidate genes in regions with no known targets have a high likelihood of also being oncogenes and TSGs.

In addition to identifying recurrent regions of chromosomal loss or gain, we considered whether recurrent breakpoints (sites where copy-number changes occur) might point to fusions between distant loci. In particular, we identified all breakpoints that recurred more than twice across the samples in our dataset, and then looked for correlations between any two of these recurrent breakpoints. We found such a correlation between recurrent breakpoints in 21q22.2 and 21q22.3 (Fig. 6). Fusions between TMPRSS2 and ERG were recently noted (35) in a large proportion of prostate cancers, and in fact all of our samples with evidence of this deletion were confirmed to have TMPRSS2-ERG fusions by

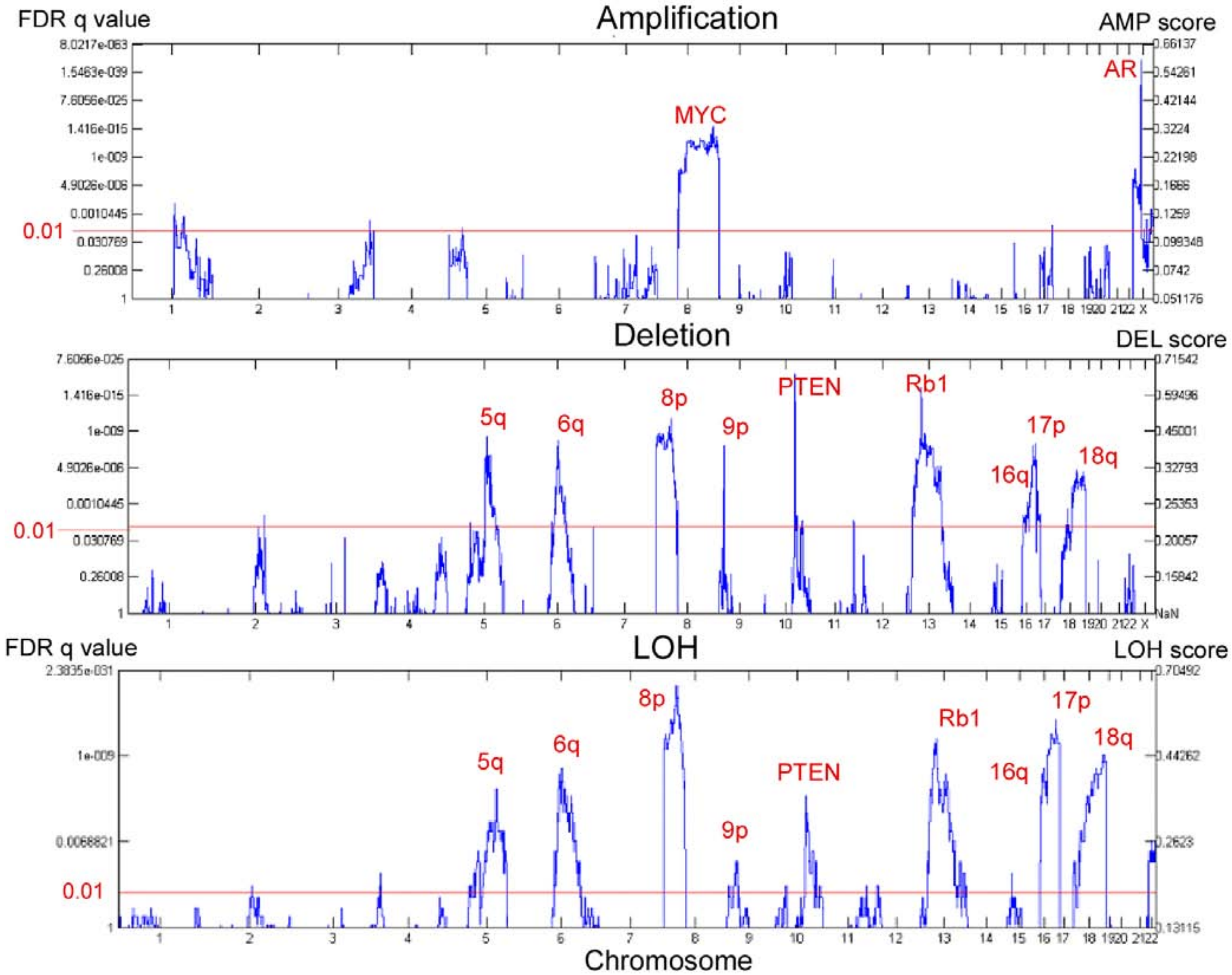


Figure 5: Regions of significant amplification, deletion, and LOH in a set of 63 prostate cancers. FDR-corrected q values (log scale, left), associated with Amp and Del scores across the genome, are displayed. Regions with a q value less than 0.01 (red line) were considered significantly altered. Among amplifications, the most significant region overlaps the androgen receptor (labeled as AR). Among deletions, the region containing PTEN scored as most significant.

assessment of the cDNA (data not shown). The presence of this interstitial deletion leading to gene fusion was further confirmed by FISH (Fig. 6) and has now been published (36).

Table 3. Locations of regions of minimal q value against known gene targets (red denotes amplified and blue deleted regions)

Location (Chro:Mb-Mb)	# of genes in region	Putative gene target	Distance from putative target
8:128.08-128.16	0	MYC	0.54 Mb
X:65.53-65.83	1	AR	Within region
10:89.35-89.52	1	PTEN	Within region
13:46.72-46.87	2	RB1	Within region
18:57.47-57.65	1	unknown	-

3. **Specific aim #3: To identify and validate candidate somatic genetic alterations differing in prevalence between localized and metastatic cancers, and develop markers for clinical association studies.**

In addition to identifying significant regions of amplification and deletion, our analysis characterizes the lesions in each sample. Thus, we can immediately identify correlations between the presence of any two or more lesions, as well as correlations between the presence of a set of lesions and phenotype. Table 4 lists the lesions that occur more frequently in metastatic than localized tumors, with q value less than 0.05. Among these, AR amplification and deletion of 9p and PTEN appear to be late events, never

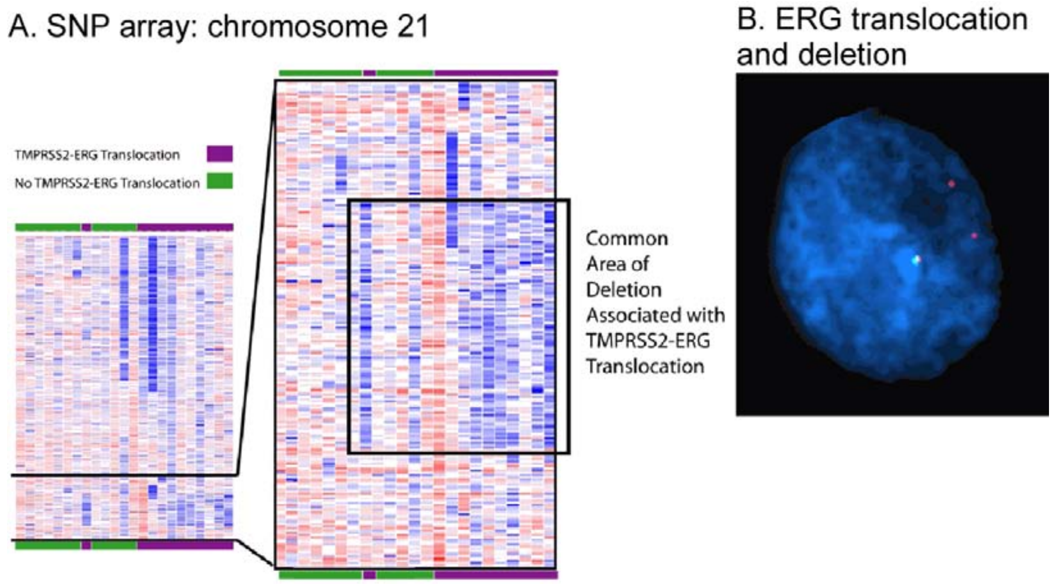

A. SNP array: chromosome 21
TMPRSS2-ERG Translocation
No TMPRSS2-ERG Translocation
Common Area of Deletion Associated with TMPRSS2-ERG Translocation
B. ERG translocation and deletion

Fig. 6: Recurrent interstitial deletion fusing TMPRSS2 and ERG. A deletion with recurrent boundaries immediately adjacent to the genes TMPRSS2 and ERG was noted across multiple samples (A), only in samples known to have undergone fusion of these genes by cDNA analysis. FISH assays (B) highlighting TMPRSS2 (red) and a region near ERG but within the interstitial deletion (green). The two probes are normally adjacent to each other, producing a yellow signal. Loss of the green signal, leading to visualization of an independent red signal, confirms deletion of the region next to ERG.

occurring in primaries. However, amplification of MYC and deletion of 18q do occur in primaries, so their enrichment in metastases suggests they may mark localized tumors that are likely to recur.

Table 4. Aberrations more commonly seen in metastases

Aberration	Relative Risk	q-value
AR amplified	Infinite	0.001
PTEN lost	Infinite	0.001
9p lost	Infinite	0.04
MYC amplified	3.7	0.04
18q lost	2.6	0.04

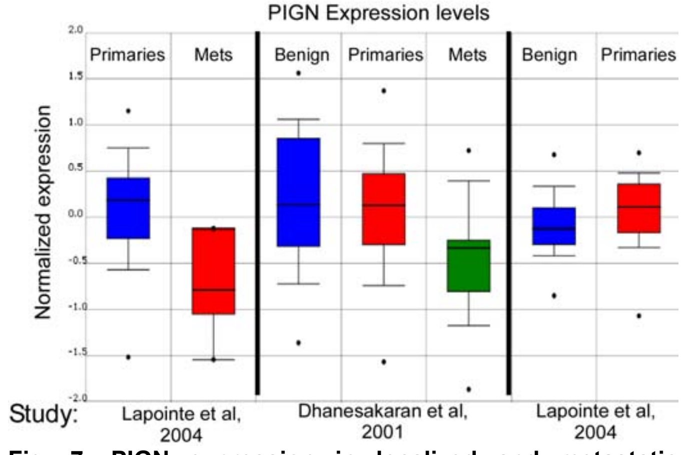


Fig. 7: PIGN expression in localized and metastatic prostate cancer, compared to benign prostate. Data from the two published studies with PIGN levels in benign prostate and localized and metastatic prostate cancer are displayed as in Oncomine (1). PIGN is one of three candidate TSG targets of the 18q deletion, which we see more commonly in metastases than primaries.

Intriguingly, although the gene targeted by 18q losses is unknown, the region with minimal q value contains only 1 gene. To assess the robustness of this finding, we excluded each sample from our dataset in turn and recalculated the boundaries of this minimal region. We found the boundaries expanded only to include a maximum of three genes. One of these is a cadherin, immediately suggesting a potential role in metastasis. Another was found, in the two expression studies of localized and metastatic cancer that assayed it (37, 38), to have lower expression in metastatic than localized tumors—although localized tumors had similar expression levels to benign tissue (Fig. 7). This is consistent with our finding that it is in a region deleted more frequently in metastases. Moreover, we note losses of the same region of 18q both by LOH and deletions (Fig. 5). Therefore, we have now validated these deletions by developing a FISH assay specifically for this region (Fig. 8).

4. Key Research Accomplishments

- Developed method for determination of LOH without paired normals that takes into account haplotype structure
- Developed methods for reducing signal-intensity errors, including systematic errors due to batch effects
- Developed methods for identifying significant regions of copy-number aberration and validated 18q losses
- Correlated several regions, including 18q losses, with progressive cancer
- Identified interstitial deletions underlying the

oncogenic fusion event between TMPRSS2 and ERG

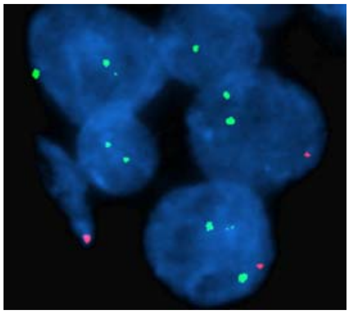


Fig. 8: FISH validation of 18q losses. The target probe (red), directed to the region of minimal q value on 18q, has a lower copy number than the reference probe (green), directed to a region infrequently altered in prostate tumors.

5. **Reportable Outcomes**

- Publication: Dutt A and **Beroukhim R**, "Single nucleotide polymorphism array analysis of cancer", Cur. Opin. Oncol., in press.
- Publication: Lee JC, Vivanco I, **Beroukhim R,** Huang JHY, Feng WL, DeBiasi RM, Yoshimoto K, King JC, Nghiemphu P, Yuza Y, Xu Q, Greulich H, Thomas RK, Paez JG, Peck TC, Linhart DJ, Glatt KA, Getz G, Onofrio R, Ziaugra L, Levine RL, Gabriel S, Kawaguchi T, O'Neill K, Khan H, Liau LM, Nelson S, Rao PN, Mischel P, Pieper RO, Cloughesy T, Leahy DJ, Sellers WR, Sawyers CL, Meyerson M, Mellinghoff IK, "EGFR activation in glioblastoma through novel missense mutations in the extracellular domain", PLoS Medicine, in press.
- Publication: Perner S*, Demichelis F*, **Beroukhim R***, Schmidt FH, Mosquera JM, Setlur S, Tchinda J, Tomlins SA, Hofer MD, Pienta KG, Kuefer R, Vessella R, Sun XW, Meyerson M, Lee C, Sellers WR, Chinnaiyan AM, Rubin MA, "TMPRSS2:ERG Fusion-Associated Deletions Provide Insight into the Heterogeneity of Prostate Cancer", Cancer Res. 2006; 66:8337.
- Publication: Garraway LA, Weir BA, Zhao X, Widlund H, **Beroukhim R**, Berger A, Rimm D, Rubin MA, Fisher DE, Meyerson ML, Sellers WR, "'Lineage addiction' in human cancer: lessons from integrated genomics", Cold Spring Harb Symp Quant Biol. 2006; 70:25.
- Publication: LaFramboise T, Weir BA, Zhao X, **Beroukhim R**, Li C, Harrington D, Sellers WR, Meyerson M, "Allele-specific amplification in cancer revealed by SNP array analysis", PLoS Comput Biol 2005;1:e65.
- Publication: **Beroukhim R***, Lin M*, Park Y, Hao K, Zhao X, Garraway LA, Fox EA, Hochberg EP, Mellinghoff IK, Hofer MD, Descazeaud A, Rubin MA, Meyerson M, Wong WH, Sellers WR, and Li C, "Inferring loss-of-heterozygosity from unpaired tumors using high-density oligonucleotide SNP Arrays", PLoS Comput Biol. 2006; 2:e41.
- Publication: Mellinghoff IK, Wang MY, Vivanco I, Haas-Kogan DA, Zhu S, Dia EQ, Lu KV, Yoshimoto K, Huang JH, Chute DJ, Riggs BL, Horvath S, Liau LM, Cavenee WK, Rao PN, **Beroukhim R**, Peck TC, Lee JC, Sellers WR, Stokoe D, Prados M, Cloughesy TF, Sawyers CL, Mischel PS, "Molecular determinants of the response of glioblastomas to EGFR kinase inhibitors", NEJM. 2005;353:2012-24.
- Publication: Garraway LA, Weir BA, Zhao X, Widlund H, **Beroukhim R**, Berger A, Rimm D, Rubin MA, Fisher DE, Meyerson ML, Sellers WR, "'Lineage Addiction' in Human Cancer: Lessons from Integrated Genomics", Cold Spring Harb Symp Quant Biol. 2005;70:1-10.
- Publication: Koochekpour S, Zhuang YJ, **Beroukhim R**, Hsieh CL, Hofer MD, Zhau HE, Hiraiwa M, Pattan DY, Ware JL, Luftig RB, Sandhoff K, Sawyers CL, Pienta KJ, Rubin MA, Vessella RL, Sellers WR, Sartor O, "Amplification and overexpression of prosaposin in prostate cancer", Genes Chromosomes Cancer. 2005; 44:351-64.
- Publication: Garraway LA, Widlund HR, Rubin MA, Getz G, Berger AJ, Ramaswamy S, **Beroukhim R**, Milner DA, Granter SR, Du J, Lee C, Wagner SN, Li C, Golub TR, Rimm DL, Meyerson ML, Fisher DE, Sellers WR, "Integrative genomic analyses identify MITF as a lineage survival oncogene amplified in malignant melanoma", Nature. 2005; 436:117-22
- Publication: Zhao X, Weir BA, LaFramboise T, Lin M, **Beroukhim R**, Garraway L, Beheshti J, Lee JC, Naoki K, Richards WG, Sugarbaker D, Chen F, Rubin MA, Janne PA, Girard L, Minna J, Christiani D, Li C, Sellers WR, Meyerson M, "Homozygous deletions and chromosome amplifications in human lung carcinomas revealed by single nucleotide polymorphism array analysis", Cancer Res. 2005; 65:5561-70.

* equal contributors

6. **Conclusions**

During the course of this award, significant progress has been made in all 3 specific aims of this grant, including determination of LOH and copy number maps using 100K SNP array data from 84 prostate tumors and use of these maps to identify chromosomal aberrations that appear to be playing a significant role in prostate cancer. Some of these regions appear to correlate with prostate cancer progression, and deletion of one of these regions has been validated by FISH. Additionally, candidate gene targets have been identified for several of these regions. In one case, deletions of 21q have been shown to comprise a mechanism leading to generation of the recently described oncogenic TMPRSS2-ERG fusion event.

7. References

1. Rhodes, D.R., Yu, J., Shanker, K., Deshpande, N., Varambally, R., Ghosh, D., Barrette, T., Pandey, A., and Chinnaiyan, A.M. 2004. ONCOMINE: a cancer microarray database and integrated data-mining platform. *Neoplasia* 6:1-6.
2. Sachidanandam, R., Weissman, D., Schmidt, S., Kakol, J., Stein, L., Marth, G., Sherry, S., Mullikin, J., Mortimore, B., Willey, D., et al. 2001. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. 409:928.
3. Cutler, D.J., Zwick, M.E., Carrasquillo, M.M., Yohn, C.T., Tobin, K.P., Kashuk, C., Mathews, D.J., Shah, N.A., Eichler, E.E., Warrington, J.A., et al. 2001. High-Throughput Variation Detection and Genotyping Using Microarrays. *Genome Res.* 11:1913-1925.
4. Matsuzaki, H., Dong, S., Loi, H., Di, X., Liu, G., Hubbell, E., Law, J., Berntsen, T., Chadha, M., Hui, H., et al. 2004. Genotyping over 100,000 SNPs on a pair of oligonucleotide arrays. *Nature Methods* 1:109.
5. Matsuzaki, H., Loi, H., Dong, S., Tsai, Y.-Y., Fang, J., Law, J., Di, X., Liu, W.-M., Yang, G., Liu, G., et al. 2004. Parallel Genotyping of Over 10,000 SNPs Using a One-Primer Assay on a High-Density Oligonucleotide Array. *Genome Res.* 14:414-425.
6. Craig, D.W., and Stephan, D.A. 2005. Applications of whole-genome high-density SNP genotyping. *Expert Review of Molecular Diagnostics* 5:159-170.
7. Huang, J., Wei, W., Zhang, J., Liu, G., Bignell, G.R., Stratton, M.R., Futreal, P.A., Wooster, R., Jones, K.W., and Shapero, M.H. 2004. Whole genome DNA copy number changes identified by high density oligonucleotide arrays. *Human Genomics* 1:287-299.
8. Irving, J.A.E., Bloodworth, L., Bown, N.P., Case, M.C., Hogarth, L.A., and Hall, A.G. 2005. Loss of Heterozygosity in Childhood Acute Lymphoblastic Leukemia Detected by Genome-Wide Microarray Single Nucleotide Polymorphism Analysis. *Cancer Res* 65:3053-3058.
9. Jacqueline A. Langdon, J.M.L.D.K.S.S.D.E.P.N.B.R.G.G.D.W.E.S.C.C. 2005. Combined genome-wide allelotyping and copy number analysis identify frequent genetic losses without copy number reduction in medulloblastoma. *Genes, Chromosomes and Cancer* 9999:NA.
10. Lieberfarb, M.E., Lin, M., Lechpammer, M., Li, C., Tanenbaum, D.M., Febbo, P.G., Wright, R.L., Shim, J., Kantoff, P.W., Loda, M., et al. 2003. Genome-wide Loss of Heterozygosity Analysis from Laser Capture Microdissected Prostate Cancer Using Single Nucleotide Polymorphic Allele (SNP) Arrays and a Novel Bioinformatics Platform dChipSNP. *Cancer Res* 63:4781-4785.
11. Lindblad-Toh, K., Tanenbaum, D.M., Daly, M.J., Winchester, E., Lui, W.-O., Villapakkam, A., Stanton, S.E., Larsson, C., Hudson, T.J., Johnson, B.E., et al. 2000. Loss-of-heterozygosity analysis of small-cell lung carcinomas using single-nucleotide polymorphism arrays. 18:1001.

12. Dumur, C.I., Dechsukhum, C., Ware, J.L., Cofield, S.S., Best, A.I.M., Wilkinson, D.S., Garrett, C.T., and Ferreira-Gonzalez, A. 2003. Genome-wide detection of LOH in prostate cancer using human SNP microarray technology. 81:260.
13. Hoque, M.O., Lee, C.-C.R., Cairns, P., Schoenberg, M., and Sidransky, D. 2003. Genome-Wide Genetic Characterization of Bladder Cancer: A Comparison of High-Density Single-Nucleotide Polymorphism Arrays and PCR-based Microsatellite Analysis. *Cancer Res* 63:2216-2222.
14. Janne, P.A., Li, C., Zhao, X., Girard, L., Chen, T.-H., Minna, J., Christiani, D.C., Johnson, B.E., and Meyerson, M. 2004. High-resolution single-nucleotide polymorphism array and clustering analysis of loss of heterozygosity in human lung cancer cell lines. *Oncogene* 23:2716-2726.
15. Mei, R., Galipeau, P.C., Prass, C., Berno, A., Ghandour, G., Patil, N., Wolff, R.K., Chee, M.S., Reid, B.J., and Lockhart, D.J. 2000. Genome-wide Detection of Allelic Imbalance Using Human SNPs and High-density DNA Arrays. *Genome Res.* 10:1126-1137.
16. Paez, J.G., Lin, M., Beroukhim, R., Lee, J.C., Zhao, X., Richter, D.J., Gabriel, S., Herman, P., Sasaki, H., Altshuler, D., et al. 2004. Genome coverage and sequence fidelity of {phi}29 polymerase-based multiple strand displacement whole genome amplification. *Nucl. Acids Res.* 32:e71-.
17. Primdahl, H., Wikman, F.P., von der Maase, H., Zhou, X.-g., Wolf, H., and Orntoft, T.F. 2002. Allelic Imbalances in Human Bladder Cancer: Genome-Wide Detection With High-Density Single-Nucleotide Polymorphism Arrays. *J Natl Cancer Inst* 94:216-223.
18. Schubert, E.L., Hsu, L., Cousens, L.A., Glogovac, J., Self, S., Reid, B.J., Rabinovitch, P.S., and Porter, P.L. 2002. Single Nucleotide Polymorphism Array Analysis of Flow-Sorted Epithelial Cells from Frozen Versus Fixed Tissues for Whole Genome Analysis of Allelic Loss in Breast Cancer. *Am J Pathol* 160:73-79.
19. Wang, Z.C., Lin, M., Wei, L.-J., Li, C., Miron, A., Lodeiro, G., Harris, L., Ramaswamy, S., Tanenbaum, D.M., Meyerson, M., et al. 2004. Loss of Heterozygosity and Its Correlation with Expression Profiles in Subclasses of Invasive Breast Cancers. *Cancer Res* 64:64-71.
20. Allinen, M., Beroukhim, R., Cai, L., Brennan, C., Lahti-Domenici, J., Huang, H., Porter, D., Hu, M., Chin, L., and Richardson, A. 2004. Molecular characterization of the tumor microenvironment in breast cancer. 6:17.
21. Beroukhim, R., Lin, M., Park, Y., Hao, K., Zhao, X., Garraway, L.A., Fox, E.A., Hochberg, E.P., Mellinghoff, I.K., Hofer, M.D., et al. 2006. Inferring loss-of-heterozygosity from unpaired tumors using high-density oligonucleotide SNP arrays. *PLoS Comput Biol* 2:e41.
22. Bignell, G.R., Huang, J., Greshock, J., Watt, S., Butler, A., West, S., Grigorova, M., Jones, K.W., Wei, W., Stratton, M.R., et al. 2004. High-Resolution Analysis of DNA Copy Number Using Oligonucleotide Microarrays. *Genome Res.* 14:287-295.
23. Zhao, X., Li, C., Paez, J.G., Chin, K., Janne, P.A., Chen, T.-H., Girard, L., Minna, J., Christiani, D., Leo, C., et al. 2004. An Integrated View of Copy Number and Allelic Alterations in the Cancer Genome Using Single Nucleotide Polymorphism Arrays. *Cancer Res* 64:3060-3071.
24. Rubin, M.A., Varambally, S., Beroukhim, R., Tomlins, S.A., Rhodes, D.R., Paris, P.L., Hofer, M.D., Storz-Schweizer, M., Kuefer, R., Fletcher, J.A., et al. 2004. Overexpression, Amplification, and Androgen Regulation of TPD52 in Prostate Cancer. *Cancer Res* 64:3814-3822.
25. Zhou, X., Mok, S.C., Chen, Z., Li, Y., and Wong, D.T.W. 2004. Concurrent analysis of loss of heterozygosity (LOH) and copy number abnormality (CNA) for oral premalignancy progression using the Affymetrix 10K SNP mapping array. *Human Genetics* 115:327.

26. Zhao, X., Weir, B.A., LaFramboise, T., Lin, M., Beroukhim, R., Garraway, L., Beheshti, J., Lee, J.C., Naoki, K., Richards, W.G., et al. 2005. Homozygous Deletions and Chromosome Amplifications in Human Lung Carcinomas Revealed by Single Nucleotide Polymorphism Array Analysis. *Cancer Res* 65:5561-5570.

27. Garraway, L.A., Widlund, H.R., Rubin, M.A., Getz, G., Berger, A.J., Ramaswamy, S., Beroukhim, R., Milner, D.A., Granter, S.R., Du, J., et al. 2005. Integrative genomic analyses identify MITF as a lineage survival oncogene amplified in malignant melanoma. *Nature* 436:117.

28. Hughes, S., Arneson, N., Done, S., and Squire, J. 2005. The use of whole genome amplification in the study of human disease. 88:173.

29. Dean, F.B., Hosono, S., Fang, L., Wu, X., Faruqi, A.F., Bray-Ward, P., Sun, Z., Zong, Q., Du, Y., Du, J., et al. 2002. Comprehensive human genome amplification using multiple displacement amplification. *PNAS* 99:5261-5266.

30. Lai, W.R., Johnson, M.D., Kucherlapati, R., and Park, P.J. 2005. Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. *Bioinformatics*:bti611.

31. Hupe, P., Stransky, N., Thiery, J.-P., Radvanyi, F., and Barillot, E. 2004. Analysis of array CGH data: from signal ratio to gain and loss of DNA regions. *Bioinformatics* 20:3413-3422.

32. Ishikawa, S., Komura, D., Tsuji, S., Nishimura, K., Yamamoto, S., Panda, B., Huang, J., Fukayama, M., Jones, K.W., and Aburatani, H. 2005. Allelic dosage analysis with genotyping microarrays. 333:1309.

33. Nannya, Y., Sanada, M., Nakazaki, K., Hosoya, N., Wang, L., Hangaishi, A., Kurokawa, M., Chiba, S., Bailey, D.K., Kennedy, G.C., et al. 2005. A Robust Algorithm for Copy Number Detection Using High-Density Oligonucleotide Single Nucleotide Polymorphism Genotyping Arrays. *Cancer Res* 65:6071-6079.

34. Benjamini, Y., and Hochberg, Y. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J Roy Stat Soc, Ser B* 57:289-300.

35. Tomlins, S.A., Rhodes, D.R., Perner, S., Dhanasekaran, S.M., Mehra, R., Sun, X.W., Varambally, S., Cao, X., Tchinda, J., Kuefer, R., et al. 2005. Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science* 310:644-648.

36. Perner, S., Demichelis, F., Beroukhim, R., Schmidt, F.H., Mosquera, J.M., Setlur, S., Tchinda, J., Tomlins, S.A., Hofer, M.D., Pienta, K.G., et al. 2006. TMPRSS2:ERG Fusion-Associated Deletions Provide Insight into the Heterogeneity of Prostate Cancer. *Cancer Res* 66:8337-8341.

37. Dhanasekaran, S.M., Barrette, T.R., Ghosh, D., Shah, R., Varambally, S., Kurachi, K., Pienta, K.J., Rubin, M.A., and Chinnaiyan, A.M. 2001. Delineation of prognostic biomarkers in prostate cancer. 412:822.

38. Lapointe, J., Li, C., Higgins, J.P., van de Rijn, M., Bair, E., Montgomery, K., Ferrari, M., Egevad, L., Rayford, W., Bergerheim, U., et al. 2004. Gene expression profiling identifies clinically relevant subtypes of prostate cancer. *PNAS* 101:811-816.

EGFR activation in glioblastoma through novel missense mutations in the extracellular domain

Jeffrey C. Lee[1,2,5], Igor Vivanco[6], Rameen Beroukhim[1,3,5], Julie H.Y. Huang[8], Whei L. Feng[1,5], Ralph M. DeBiasi[1,5], Koji Yoshimoto[11], Jennifer C. King[7], Phioanh Nghiemphu[9], Yuki Yuza[1], Qing Xu[1,3], Heidi Greulich[1,3,5], Roman K. Thomas[1,5], J. Guillermo Paez[1,5], Timothy C. Peck[1,5], David J. Linhart[1,5], Karen A. Glatt[1], Gad Getz[5], Robert Onofrio[5], Liuda Ziaugra[5], Ross L. Levine[1,4], Stacey Gabriel[5], Tomohiro Kawaguchi[13], Keith O'Neill[5], Haumith Khan[10], Linda M. Liau[10], Stan Nelson[8], P. Nagesh Rao[11], Paul Mischel[11], Russell O. Pieper[13], Tim Cloughesy[9], Daniel J. Leahy[14], William R. Sellers[1,3,5], Charles L. Sawyers[6,7,12], Matthew Meyerson[1,2,5]#, and Ingo K. Mellinghoff[6,7]#

[1]Department of Medical Oncology and Center for Cancer Genome Discovery, Dana-Farber Cancer Institute, Departments of [2]Pathology and [3]Medicine, Harvard Medical School, [4]Department of Medicine, Brigham and Women's Hospital, Boston, MA 02115, USA. [5]Broad Institute of Harvard and MIT, Cambridge, MA 02141, USA. [6] Departments of Molecular & Medical Pharmacology, [7]Medicine, [8]Genetics, [9]Neurology, [10]Neurosurgery, [11]Pathology and Laboratory Medicine, David Geffen School of Medicine, and [12]Howard Hughes Medical Institute, University of California, Los Angeles, CA 90095, USA. [13]Department of Neurosurgery, University of California, San Francisco, CA 94115, USA. [14]Department of Biophysics and Biophysical Chemistry and Howard Hughes Medical Institute, The Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA.

# corresponding authors

ABSTRACT

Background

Protein tyrosine kinases are important regulators of cellular homeostasis with tightly controlled catalytic activity. Mutations in kinase-encoding genes can relieve the autoinhibitory constraints on kinase activity, can promote malignant transformation, and appear to be a major determinant of response to kinase inhibitor therapy [1,2].

Methods and Findings

Encouraged by the promising clinical activity of epidermal growth factor receptor (EGFR) kinase inhibitors in glioblastoma [3-5], we have sequenced the complete EGFR coding sequence in glioma tumor samples and cell lines. We identified novel missense mutations in the extracellular domain of *EGFR* in 18/132 (13.6 %) glioblastomas and 1/8 (12.5 %) glioblastoma cell lines. These *EGFR* mutations were associated with increased *EGFR* gene dosage and conferred anchorage-independent growth and tumorigenicity to NIH-3T3 cells. Cells transformed by expression of these *EGFR* mutants were sensitive to small-molecule EGFR kinase inhibitors.

Conclusion

Our results suggest extracellular missense mutations as a novel mechanism for oncogenic EGFR activation and may help identify patients who can benefit from EGFR kinase inhibitors for glioblastoma.

INTRODUCTION

The epidermal growth factor receptor (EGFR) is a receptor tyrosine kinase that regulates fundamental processes of cell growth and differentiation. Deletion of the *EGFR* gene is embryonically lethal in mice and increased EGFR signaling has been linked to a variety of human malignancies. Mechanisms for oncogenic conversion of EGFR in cancer include *EGFR* gene amplification, structural rearrangements of the receptor, overexpression of EGF-family ligands by tumor cells and/or surrounding stroma, and – as recently shown in lung cancer – activating mutations in the *EGFR* kinase domain [6].

The evidence for a role of EGFR in oncogenesis is particularly compelling in glioblastoma, the most aggressive human brain tumor [7]. About 40 % of glioblastomas show amplification of the EGFR gene locus [8] and about half of these tumors express a mutant receptor (EGFRvIII) that is constitutively active due to an in-frame truncation within the extracellular ligand-binding domain [9-11]. Perhaps the strongest evidence for a role of EGFR in the biology of glioblastoma stems from clinical trials where 15-20 % of glioblastoma patients experienced significant tumor regression in response to small molecule EGFR kinase inhibitors [4,5]. Our recent data indicates that expression of EGFRvIII in the context of an intact PTEN pathway is associated with these clinical responses [4].

To explore the possibility that *EGFR* might be the target of oncogenic mutations outside the kinase domain, we sequenced the entire EGFR coding region in a panel of 151 glioma tumors and cell lines. We report the discovery of novel extracellular missense mutations in *EGFR* in 13.6 % (18/132) of human glioma samples. These mutations are oncogenic in cellular transformation assays and sensitize transformed cells to the antiproliferative effects of small molecule EGFR kinase inhibitors. Only one of our clinical glioma samples (< 1 %) harbored a mutation within the *EGFR* kinase domain, supporting the recent conclusion from other groups that *EGFR* kinase domain mutations appear to be a rare event in this disease [12,13] [14]. Interestingly, further examination of 119 primary lung tumors from our previous study [15] showed a distinctly different distribution of EGFR mutations with *EGFR* kinase domain mutations in 13.4 % (16/119) of cases and no evidence for extracellular *EGFR* missense mutations. These results indicate that mechanisms of oncogenic kinase conversion may differ considerably between tumor types and warrant an extension of current kinome resequencing efforts beyond the kinase domain.

METHODS

<u>DNA samples.</u> Genomic DNA was extracted from eight glioblastoma cell lines (A172, SF268, SF295, SF539, T98G, U87, U118, and U251) and 143 fresh frozen glioma samples. The clinical glioma samples comprised of glioblastomas (n = 132), WHO grade III anaplastic astrocytomas (n = 3), grade III mixed gliomas (n = 4), and grade III oligodendrogliomas (n = 4). Germline genomic DNA was extracted from peripheral blood samples. To confirm the match between germline and tumor DNA for each patient, we performed mass spectrometric genotyping of 24 single-nucleotide polymorphism (SNP) loci. These loci included 23 SNP loci represented on both 50K Xba and Hind arrays (Affymetrix, Santa Clara, CA) and one AmelXY locus for sex determination *(Supplementary Table S1)*. Collection and analysis of all clinical samples was approved by the UCLA Institutional Review Board.

<u>Reagents.</u> Retroviral constructs for EGFR and EGFRvIII were generously provided by Dr. David Riese 2nd and Dr. Webster Cavenee. Erlotinib was purchased from WuXi Pharmatech (Shanghai, China). The following antibodies were used in this study: anti-EGFR, anti-phospho-

Y1068 EGFR, anti-phospho-Y845-EGFR, anti-phosphoinositide 3-kinase (PI3K) p85. (Cell Signaling Technology, Beverly, MA); anti-phosphotyrosine 4G10 (Upstate Biotechnologies, Waltham, MA); anti-actin, anti-ERK1/2, and anti-P-ERK1/2 (Santa Cruz Biotechnology, Santa Cruz, CA).

Sequencing and Mass Spectrometric Genotyping. PCR reactions for each exon and flanking intronic sequences contained 5 ng of genomic DNA, 1X HotStar Buffer, 0.8 mM dNTPs, 1 mM $MgCl_2$, 0.2U HotStar Enzyme (Qiagen, Valencia, CA), and 0.2 µM forward and reverse primers in a 6 or 10 µL reaction volume. PCR cycling parameters were: one cycle of 95°C for 15 min, 35 cycles of 95°C for 20 seconds, 60°C for 30 seconds and 72°C for 1 minute, followed by one cycle of 72°C for 3 minutes. The resulting PCR products were sequenced using bi-directional dye-terminator fluorescent sequencing with universal M13 primers. Sequencing fragments were detected via capillary electrophoresis using ABI Prism 3730 DNA Analyzer (Applied Biosystems, Foster City, CA). PCR and sequencing were performed at Agencourt Bioscience Corporation (Beverly, MA) or at the Broad Institute of Harvard and MIT (Cambridge, MA). Forward (F) and reverse (R) chromatograms were analyzed in batch with Mutation Surveyor 2.51 (SoftGenetics, State College, PA), followed by manual review. A minimum of 21 of 28 (75%) EGFR exon sequence coverage was accomplished for 151 samples. An exon for each individual sample was considered covered if 90% of the sequence trace within the exon had a phred quality score of 30 or greater, a signal to background noise ration of 15% or less, and signal intensity greater than 1/4 of the signal intensity of the sequencing plate. High quality sequence variations found in one or both directions were scored as candidate mutations. Exons harboring candidate mutations were reamplified from the original DNA sample and resequenced. For mass spectrometric genotyping, PCR and extension primers *(Supplementary Table S2)* were designed using SpectroDESIGNER software (Sequenom, San Diego, CA). Unincorporated nucleotides from PCR reactions were dephosphorylated with shrimp alkaline phosphatase (Amersham) followed by primer extension with ThermoSequence polymerase (Amersham, Piscataway, NJ). Primer extension reactions were loaded onto SpectroCHIPs (Sequenom) and analyzed using a MALDI-TOF (matrix-assisted laser desorption/ionization time-of-flight) mass spectrometer (SpectroREADER, Sequenom) [16]. Mass spectra were processed with SpectroTYPER (Sequenom) to determine genotypes based on peaks intensities corresponding to the expected extension products.

Affymetrix 100k SNP arrays. Genomic DNA was processed and hybridized following the guidelines of the manufacturer (Affymetrix) and arrays were scanned with a GeneChip Scanner 3000. Genotyping calls and signal quantification were obtained using GeneChip Operating System 1.1.1 and Affymetrix Genotyping Tools 2.0 software. Data were normalized at the probe level to a baseline array with median signal intensity using invariant set normalization. After normalization, the signal values for each SNP in each array were obtained with a model-based (PM/MM) method [17]. Signal intensities at each probe locus were compared with a set of normal reference samples representing 36 ethnically matched individuals to generate log2 ratios. Log2 ratios were smoothed using the break-point analysis method in the R package GLAD (Gain and Loss Analysis of DNA) [18]. Regions were considered amplified if their smoothed log2 ratio exceeded 0.3 (half the variation seen with a single-copy gain).

Fluorescence-in-situ-hybridization (FISH). Dual-probe fluorescence in situ hybridization was performed on paraffin-embedded sections with locus-specific probes for *EGFR* and the centromere of chromosome 7 as previously described [4].

Determination of *EGFRvIII* expression. RNA was extracted from fresh frozen tumor samples and *EGFRvIII* expression determined by two independent RT-PCR assays for each sample. Primer pairs included: #1F 5' CTT CGG GGA GCA GCG ATG CGA C 3', #1R 5' ACC AAT ACC TAT TCC GTT ACA C 3', #2F GAGCTCTTCGGGGAGCAG, and #2R GTGATCTGTCACCACATAATTACCTTTCTT. EGFRvIII expression was also examined by immunohistochemistry and/or immunoblotting depending on the availability of tissue samples.

Quantification of mutant EGFR alleles. The abundance of missense and wildtype *EGFR* alleles in tumor DNA samples was determined by PCR-cloning and sequencing of respective *EGFR* exons (see Table S2). PCR products were ligated into PCR2.1-Topo vectors (Invitrogen) and transformed into E.coli. After transformation, bacteria were plated onto selection plates and grown overnight. 65-94 colonies were isolated for each DNA sample using a colony picking robot (QPix2, Genetix Limited), grown overnight, and bidirectionally sequenced at the Broad Institute. Sequence traces were analyzed using Mutation Surveyor software (SoftGenetic Inc.).

EGFR Expression Constructs. Retroviral EGFR expression constructs containing puromycin (pBabe-puro-*EGFR*) [19] or neomycin-resistance genes (pLXSN-neo-*EGFR*) [20] were used for site-directed mutagenesis using the Quick-Change Mutagenesis XL kit (Stratagene, La Jolla, CA). pBabe-Puro-based viral stocks were generated by transfecting the Phoenix 293T packaging cell line (Orbigen, San Diego, CA) with the pBabe-Puro retroviral constructs using Lipofectamine 2000 (Invitrogen, Carlsbad, CA). pLXSN-Neo-based viral stocks were generated by transfecting the human amphotrophic 293-T cell line with pLXSN-Neo retroviral constructs using Lipofectamine 2000 (Invitrogen). Supernatants were collected 24-48 hours post-transfection, filtered (0.45 µM), and used to infect NIH-3T3 cells, Ba/F3 cells, and human astrocytes.

Expression of *EGFR* alleles in NIH-3T3 cells. Cells cultured in DMEM supplemented with 10% calf serum were infected with pBabe-Puro-based viral stock in the presence of polybrene. Two days after infection, cells were selected in puromycin (2 μg/ml) for 3 days. Pooled NIH-3T3 cells stably expressing respective EGFR alleles at comparable EGFR protein levels were examined for their ability to induce colony formation in soft agar and tumor growth in nude mice. For soft agar assays, $1x10^5$ NIH-3T3 cells were suspended in a top layer of DMEM supplemented with 10% calf serum and 0.4% Select Agar (Gibco/Invitrogen) and plated on a bottom layer of DMEM supplemented with 10% calf serum and 0.5% Select Agar. EGF (10 ng/ml) was added to the top agar where indicated in the figure legend. Pictures of colonies were taken 2-3 weeks after plating. Colonies were counted from ten random images (40x magnification) taken from each well. Colonies were counted from three replicate wells with the average number represented. In-vivo tumorigenicity assays were performed in three mice (two injections/mice) for each cell line. For each injection, 2×10^6 cells were injected subcutaneously into each nude mice (Taconic, Germantown, NY) and three-dimensional tumor volumes calculated 3-4 weeks following injection.

Expression of *EGFR* alleles in Ba/F3 cells. Murine Ba/F3 pro-B lymphocytes [21] were cultured in RPMI 1640 (Cellgro, Herndon, CA) supplemented with 10% FCS, 100 units/mL penicillin and 100 μg/mL streptomycin, 1% L-glutamine, and 10% WEHI3B conditioned media. To derive Ba/F3 subclones stably expressing various EGFR alleles, Ba/F3 cells were spinfected with pLXSN-neo-*EGFR*-based (Figure 3A) or pBabe-puro-*EGFR*- based (in Figure 4D) viral supernatants and spinfection repeated after 48 hours. Cells were selected for neomycin or puromycin resistance and maintained in the presence of IL-3. IL-3 independent subclones were derived through prolonged passage in IL-3 depleted media. To determine sensitivity to erlotinib, 1 x 10^3 cells were seeded in 96-well flat bottomed plates with the indicated concentrations of erlotinib. Cell proliferation was assessed 48 hours post-plating using the WST-1 assay (Roche, Indianapolis, IN). Each data point represents the median of six replicate wells for each Ba/F3 subclone and erlotinib concentration.

Expression of *EGFR* alleles in human astrocytes. Viral supernatants (pLXSN-neo-EGFR) were used to infect immortalized human astrocytes expressing the catalytic subunit of the telomerase holoenzyme and human papillomavirus 16 E6/E7 [22]. Astrocytes were then selected in G418 (Invitrogen) for approximately 10 days.

RESULTS

Missense Mutations in Glioblastoma Cluster in the Extracellular Domain of *EGFR*

Encouraged by the recent success in identifying oncogenic kinase mutations through resequencing of kinase-encoding genes [15,23,24], we sequenced the entire coding sequence of EGFR in 143 human glioma samples and eight glioblastoma cell lines. Analysis of the initial Sanger sequencing results in these 151 samples revealed several novel sequence variations in the coding region of the EGF receptor. To validate these candidate mutations via a complementary method, all DNA samples were reexamined using allele-specific genotyping by matrix-assisted laser desorption ionization time-of-flight (MALDI-TOF) mass spectrometry. In all, we identified *EGFR* missense mutations in 19/132 (14.4 %) glioblastomas, 1/8 (12.5 %) glioblastoma cell lines, and none (0/11) in lower grade gliomas. Remarkably, only one tumor sample harbored a missense mutation in the *EGFR* kinase domain (L861Q), the location of EGFR mutations in lung cancer, whereas the remainder of the *EGFR* mutations (18/132 glioblastomas) were located in the extracellular ligand-binding (I, III) or cysteine-rich (II, IV) domains of the receptor *(Figure 1A).* Two evolutionarily highly conserved amino acid residues *(Supplementary Figure S1)* were affected by mutations in five samples each (R108 and A289). Examination of peripheral blood DNA, matched to the tumor DNA by genotyping of 24 SNP loci, showed that eight of the twelve distinct missense mutations were unambiguously somatic and one mutation (E330K) was germline. Three additional missense mutations (A289D, A289T, and R324L) were found in tumors for which no normal tissue was available *(Table 1).* None of the missense mutations were detected in germline DNA from 270 normal control individuals.

To define which fraction of the *EGFR* pool represented the mutant allele in gliomas with *EGFR* missense mutations, we employed a PCR-cloning strategy previously used by our laboratories for mutation detection in clinical samples [25]. The mutant EGFR allele represented 30- 98 % of the receptor pool in two thirds of all examined cases (10/16) and > 50 % in at least one tumor representing the most common genotypes R108K, T263P, A289V, and G598V (Table 1). Lower abundance of the mutant *EGFR* allele in other samples might be due to contaminating stromal tissue since genomic DNA was extracted from frozen tumor aliquots without prior microdissection.

We also genotyped genomic DNA from 119 primary lung tumors for the presence of the *EGFR* ectodomain mutations. While 16/119 (13.4 %) of these lung tumor samples harbored mutations in the *EGFR* kinase domain, we found none of the glioma-related *EGFR* ectodomain mutations in this sample set.

***EGFR* Ectodomain Mutations are Associated with Increased *EGFR* Gene Dose**

Since *EGFR* is amplified in about 40% of human glioblastomas [8], we determined the relationship between *EGFR* missense mutation and *EGFR* gene dose in our tumor samples.

10/17 (58.8 %) tumors with *EGFR* missense mutations showed evidence for *EGFR* amplification by fluorescence-in-situ hybridization (FISH) and/or Affymetrix 100K single nucleotide polymorphism (SNP) genotyping arrays *(Figure 1B)(Table 1)*. This distribution suggested that *EGFR* missense mutations are associated with *EGFR* amplification and raised the question whether *EGFR* missense mutations in glioblastoma co-occur with or are mutually exclusive with the *EGFRvIII* mutation, which is almost exclusively found in glioblastomas with increased gene dosage [26]. Using at least two independent assays for EGFRvIII determination, we identified EGFRvIII in 13/46 (28.3 %) of gliomas without *EGFR* missense mutation and 1/16 (6.3 %) tumors with *EGFR* missense mutation *(Figure 1C)(Table 1)*; note that this tumor showed vastly lower levels of *EGFRvIII (Figure 1C, lane 12)*. These findings suggest that *EGFR* ectodomain mutations occur independently of *EGFRvIII* in glioblastoma and provide an alternative mechanism for EGFR activation in this disease.

EGFR Ectodomain Mutants are Oncogenic

To test the oncogenicity of the glioma-related *EGFR* missense mutations, we transduced NIH-3T3 fibroblasts with retroviruses encoding either wild-type *EGFR* or selected *EGFR* missense mutants (R108K, T263P, A289V, G598V, L861Q). Ectopic expression of all *EGFR* mutants examined in NIH-3T3 cells conferred anchorage-independent colony formation in soft agar *(Figure 2A)*. In contrast, expression of wild-type *EGFR* induced a transformed phenotype only in the presence of exogenous EGF, as previously reported [27,28].

To further analyze the oncogenic potential of the *EGFR* mutants, NIH-3T3 subclones stably expressing mutant receptors (R108K, T263P, A289V, G598V, L861Q) were inoculated subcutaneously into nude mice. NIH-3T3 cells infected with empty vector or wildtype *EGFR*–

expressing virus did not yield any measurable tumors within the four-week observation period. In contrast, NIH-3T3 cells expressing each of the tested *EGFR* missense mutants produced large tumors at the inoculation site in all mice within three to four weeks (*Figure 2B*).

EGFR ectodomain mutants are basally phosphorylated and are responsive to ligand.

Signal transduction through EGFR is determined by its basal catalytic activity, receptor activation by ligand, and signal termination through intracellular compartmentalization of the receptor-ligand complex, receptor dephosphorylation, and degradation [29]. To explore the biochemical basis for the gain of function observed with *EGFR* ectodomain mutants, we first examined the basal catalytic activity of A289V-EGFR in transiently transfected 293T cells using EGFR autophosphorylation as a readout for receptor activation. EGFR autophosphorylation was determined by measuring the total phosphotyrosine content of the immunoprecipitated receptor *(Figure 3A, left panel)* and by immunoblotting of whole cell lysates with phosphosite-specific anti-EGFR antibodies *(Figure 3A,right panel)*. Compared to wildtype EGFR, the ectodomain mutant A289V-EGFR showed a marked increase in receptor autophosphorylation in the absence of ligand or serum. We subsequently examined a more extensive panel of EGFR missense mutants (T263P, A289V, G598V, L861Q) in immortalized human astrocytes [22] stably transduced with these receptors. Compared to astrocytes overexpressing wildtype EGFR, sublines expressing EGFR missense mutants showed an increased phosphotyrosine content of EGFR and several other unidentified proteins under serum-free conditions *(Figure 3B)*.

We also expressed selected *EGFR* mutants (R108K, T263P, A289V, G598V, L861Q) in murine hematopoietic cells (Ba/F3 cells) which do not express any EGF receptor family members [20] but otherwise retain functional properties of the EGF-signaling pathway [30,31]

[32] [33]. Consistent with our findings in 293T cells and astrocytes, all examined EGFR ectodomain mutants showed increased tyrosine phosphorylation under serum-starved conditions and were responsive to exogenous EGF *(Figure 3C)*. We also noted that EGF stimulation led to a more pronounced drop of EGFR levels in Ba/F3 cells expressing wildtype EGFR than in subclones expressing EGFR ectodomain mutants *(Figure 3C)*, reminiscent of the impaired ligand-induced receptor downregulation reported for selected EGFR kinase domain mutants [33].

Sensitivity of EGFR Ectodomain Mutants to EGFR Kinase Inhibitors

The presence of identical missense mutations in multiple patient samples and their oncogenicity in standard transformation assays suggest that these mutants play a role in gliomagenesis. It also raises the question whether these mutations might sensitize transformed cells to EGFR kinase inhibitors. Ba/F3 cells provide a unique model system to examine kinase inhibitor sensitivity [34-37] because stable expression of oncogenic kinases in these cells can relieve them from their intrinsic dependence on interleukin-3 (IL-3) for survival [21] [38]. As expected from our results in NIH-3T3 cells, expression of the tested EGFR missense mutants but not wildtype EGFR was able to relieve Ba/F3 cells from IL-3 dependence. Addition of the EGFR kinase inhibitor erlotinib to the media had little or no effect on the viability of parental Ba/F3 cells growing in the presence of IL-3 or on Ba/F3 cells expressing the drug-resistant EGFR double-mutant L858R/T790M-EGFR. However, erlotinib did induce dose-dependent cell death in Ba/F3 subclones expressing the *EGFR* ectodomain mutants (missense and vIII truncation) or *EGFR* kinase domain mutants (L858R and L861Q)*(Figure 4A)*. Of note, erlotinib-induced cell death of Ba/F3 cells expressing *EGFR* ectodomain mutants occurred at IC-50 values of 50-150 nM, drug concentrations that are well below the concentrations achieved in human plasma [39]. These data suggest that *EGFR* missense mutants sensitize transformed cells to EGFR kinase inhibitors similar to EGFRvIII or lung cancer-related kinase domain mutants, both of which have been associated with clinical responses to EGFR kinase inhibitor therapy [4,15,40,41].

We recently reported the results of a glioblastoma clinical trial with EGFR kinase inhibitors which associated clinical responses to the co-expression of EGFRvIII and PTEN [4]. To investigate whether clinical responses might also be linked to the presence of EGFR ectodomain mutants, we reexamined all available tumor DNA samples from this clinical trial. We identified the ectodomain mutant R108K-*EGFR* in one of seven (14%) gliomas that responded to erlotinib. This tumor, however, also expressed *EGFR*vIII, raising the possibility of independent clones arising from a common progenitor with *EGFR* amplification. We also identified the R108K *EGFR* mutation in 1/15 (7 %) gliomas that failed EGFR kinase inhibitor therapy, but loss of PTEN in this tumor provides a potential explanation for treatment failure *(Supplementary Table S3)*. Larger clinical trials are required to ascertain the contribution of EGFR missense mutants to EGFR kinase inhibitor response in glioblastoma.

DISCUSSION

We have identified novel oncogenic missense mutations in the ectodomain of EGFR in glioma. The association of these mutations with increased *EGFR* gene dosage raises the question whether similar ectodomain missense mutations might exist in other malignancies with *EGFR* amplification or polysomy of chromosome 7. More broadly, our results suggest that ectodomain missense mutations in other tyrosine kinase genes may be transforming events in multiple cancers, and argue for an extension of current kinase gene resequencing efforts beyond the kinase domains [42,43].

The ligand-independent basal phosphorylation of the EGFR missense mutants in our study is consistent with their ability to confer NIH3T3 cells with the ability to grow in soft agar in the absence of exogenous EGF. Whether all EGFR ectodomain mutants share a common mechanism of oncogenic receptor conversion warrants further study. A common mechanism is suggested by the structural observation fact that many of the resulting amino-acid substitutions map to interdomain interfaces. R108K and A289V/D/T occur at the domain I/II interface, P569L and G598V occur at the domain II/IV contact, and T263P occurs in domain II just prior to the extended loop that contacts domain IV *(Figure 4b).* Differences in constitutive receptor activity (G598V>A289V>T263P), on the other hand, point toward alternative mechanisms of oncogenic receptor conversion.

Three of the EGFR missense mutations (P596L, G598V, A289V) were previously observed in smaller cohorts of glioblastoma tumors [26] [44]. The identification of additional ectodomain mutations in our study might have been facilitated by the large number of tumors, near complete coverage of the EGFR coding sequence, and use of MALDI-TOF mass spec genotyping in addition to Sanger sequencing. Since most of the patients in our study were of Caucasian descent, we were unable to establish whether the prevalence of EGFR ectodomain mutations in glioblastoma might be affected by ethnicity as has been shown for EGFR kinase domain mutations. The distribution of EGFR missense mutations in glioblastoma (largely extracellular) and lung cancer (exclusively kinase domain) suggests fundamental differences in oncogenic EGFR signaling between these two tumor types. Importantly, however, both classes of mutants – as well as EGFRvIII - appear to sensitize transformed cells to EGFR kinase inhibitors in a preclinical model system that has been predictive of clinical responses [35] [45]. Based on the experience with kinase inhibitors for chronic myeloid leukemia, the development of sensitive methodologies to monitor the EGFR pool before and during therapy [46] will constitute an important step to advance the current use of EGFR kinase inhibitors for cancer.

REFERENCES

1. Schlessinger J (2003) Signal transduction. Autoinhibition control. Science 300: 750-752.

2. Sawyers C (2004) Targeted cancer therapy. Nature 432: 294-297.

3. Rich JN, Reardon DA, Peery T, Dowell JM, Quinn JA, et al. (2004) Phase II trial of gefitinib in recurrent glioblastoma. J Clin Oncol 22: 133-142.

4. Mellinghoff IK, Wang MY, Vivanco I, Haas-Kogan DA, Zhu S, et al. (2005) Molecular determinants of the response of glioblastomas to EGFR kinase inhibitors. N Engl J Med 353: 2012-2024.

5. Haas-Kogan DA, Prados MD, Tihan T, Eberhard DA, Jelluma N, et al. (2005) Epidermal growth factor receptor, protein kinase B/Akt, and glioma response to erlotinib. J Natl Cancer Inst 97: 880-887.

6. Hynes NE, Lane HA (2005) ERBB receptors and cancer: the complexity of targeted inhibitors. Nat Rev Cancer 5: 341-354.

7. Kleihues P, Louis DN, Scheithauer BW, Rorke LB, Reifenberger G, et al. (2002) The WHO classification of tumors of the nervous system. J Neuropathol Exp Neurol 61: 215-225; discussion 226-219.

8. Libermann TA, Nusbaum HR, Razon N, Kris R, Lax I, et al. (1985) Amplification, enhanced expression and possible rearrangement of EGF receptor gene in primary human brain tumours of glial origin. Nature 313: 144-147.

9. Yamazaki H, Fukui Y, Ueyama Y, Tamaoki N, Kawamoto T, et al. (1988) Amplification of the structurally and functionally altered epidermal growth factor receptor gene (c-erbB) in human brain tumors. Mol Cell Biol 8: 1816-1820.

10. Wong AJ, Ruppert JM, Bigner SH, Grzeschik CH, Humphrey PA, et al. (1992) Structural alterations of the epidermal growth factor receptor gene in human gliomas. Proc Natl Acad Sci U S A 89: 2965-2969.

11. Ekstrand AJ, Sugawa N, James CD, Collins VP (1992) Amplified and rearranged epidermal growth factor receptor genes in human glioblastomas reveal deletions of sequences encoding portions of the N- and/or C-terminal tails. Proc Natl Acad Sci U S A 89: 4309-4313.

12. Rich JN, Rasheed BK, Yan H (2004) EGFR mutations and sensitivity to gefitinib. N Engl J Med 351: 1260-1261; author reply 1260-1261.

13. Barber TD, Vogelstein B, Kinzler KW, Velculescu VE (2004) Somatic mutations of EGFR in colorectal cancers and glioblastomas. N Engl J Med 351: 2883.

14. Marie Y, Carpentier AF, Omuro AM, Sanson M, Thillet J, et al. (2005) EGFR tyrosine kinase domain mutations in human gliomas. Neurology 64: 1444-1445.

15. Paez JG, Janne PA, Lee JC, Tracy S, Greulich H, et al. (2004) EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy. Science 304: 1497-1500.

16. Ross P, Hall L, Smirnov I, Haff L (1998) High level multiplex genotyping by MALDI-TOF mass spectrometry. Nat Biotechnol 16: 1347-1351.

17. Li C, Wong WH (2001) Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. Proc Natl Acad Sci U S A 98: 31-36.

18. Hupe P, Stransky N, Thiery JP, Radvanyi F, Barillot E (2004) Analysis of array CGH data: from signal ratio to gain and loss of DNA regions. Bioinformatics 20: 3413-3422.

19. Greulich H, Chen TH, Feng W, Janne PA, Alvarez JV, et al. (2005) Oncogenic Transformation by Inhibitor-Sensitive and -Resistant EGFR Mutants. PLoS Med 2: e313.

20. Riese DJ, 2nd, van Raaij TM, Plowman GD, Andrews GC, Stern DF (1995) The cellular response to neuregulins is governed by complex interactions of the erbB receptor family. Mol Cell Biol 15: 5770-5776.

21. Palacios R, Steinmetz M (1985) Il-3-dependent mouse clones that express B-220 surface antigen, contain Ig genes in germ-line configuration, and generate B lymphocytes in vivo. Cell 41: 727-734.

22. Sonoda Y, Ozawa T, Hirose Y, Aldape KD, McMahon M, et al. (2001) Formation of intracranial tumors by genetically modified human astrocytes defines four pathways critical in the development of human anaplastic astrocytoma. Cancer Res 61: 4956-4960.

23. Davies H, Bignell GR, Cox C, Stephens P, Edkins S, et al. (2002) Mutations of the BRAF gene in human cancer. Nature 417: 949-954.

24. Samuels Y, Wang Z, Bardelli A, Silliman N, Ptak J, et al. (2004) High frequency of mutations of the PIK3CA gene in human cancers. Science 304: 554.

25. Gorre ME, Mohammed M, Ellwood K, Hsu N, Paquette R, et al. (2001) Clinical resistance to STI-571 cancer therapy caused by BCR-ABL gene mutation or amplification. Science 293: 876-880.

26. Frederick L, Wang XY, Eley G, James CD (2000) Diversity and frequency of epidermal growth factor receptor mutations in human glioblastomas. Cancer Res 60: 1383-1387.

27. Velu TJ, Beguinot L, Vass WC, Willingham MC, Merlino GT, et al. (1987) Epidermal-growth-factor-dependent transformation by a human EGF receptor proto-oncogene. Science 238: 1408-1410.

28. Di Fiore PP, Pierce JH, Fleming TP, Hazan R, Ullrich A, et al. (1987) Overexpression of the human EGF receptor confers an EGF-dependent transformed phenotype to NIH 3T3 cells. Cell 51: 1063-1070.

29. Wiley HS (2003) Trafficking of the ErbB receptors and its influence on signaling. Exp Cell Res 284: 78-88.

30. Pierce JH, Ruggiero M, Fleming TP, Di Fiore PP, Greenberger JS, et al. (1988) Signal transduction through the EGF receptor transfected in IL-3-dependent hematopoietic cells. Science 239: 628-631.

31. Collins MK, Downward J, Miyajima A, Maruyama K, Arai K, et al. (1988) Transfer of functional EGF receptors to an IL3-dependent cell line. J Cell Physiol 137: 293-298.

32. Walker F, Hibbs ML, Zhang HH, Gonez LJ, Burgess AW (1998) Biochemical characterization of mutant EGF receptors expressed in the hemopoietic cell line BaF/3. Growth Factors 16: 53-67.

33. Yang S, Qu S, Perez-Tores M, Sawai A, Rosen N, et al. (2006) Association with HSP90 inhibits Cbl-mediated down-regulation of mutant epidermal growth factor receptors. Cancer Res 66: 6990-6997.

34. Cools J, Stover EH, Boulton CL, Gotlib J, Legare RD, et al. (2003) PKC412 overcomes resistance to imatinib in a murine model of FIP1L1-PDGFRalpha-induced myeloproliferative disease. Cancer Cell 3: 459-469.

35. Shah NP, Tran C, Lee FY, Chen P, Norris D, et al. (2004) Overriding imatinib resistance with a novel ABL kinase inhibitor. Science 305: 399-401.

36. Growney JD, Clark JJ, Adelsperger J, Stone R, Fabbro D, et al. (2005) Activation mutations of human c-KIT resistant to imatinib mesylate are sensitive to the tyrosine kinase inhibitor PKC412. Blood 106: 721-724.

37. Jiang J, Greulich H, Janne PA, Sellers WR, Meyerson M, et al. (2005) Epidermal growth factor-independent transformation of Ba/F3 cells with cancer-derived epidermal growth factor receptor mutants induces gefitinib-sensitive cell cycle progression. Cancer Res 65: 8968-8974.

38. Daley GQ, Baltimore D (1988) Transformation of an interleukin 3-dependent hematopoietic cell line by the chronic myelogenous leukemia-specific P210bcr/abl protein. Proc Natl Acad Sci U S A 85: 9312-9316.

39. Baselga J, Arteaga CL (2005) Critical update and emerging trends in epidermal growth factor receptor targeting in cancer. J Clin Oncol 23: 2445-2459.

40. Lynch TJ, Bell DW, Sordella R, Gurubhagavatula S, Okimoto RA, et al. (2004) Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib. N Engl J Med 350: 2129-2139.

41. Pao W, Miller V, Zakowski M, Doherty J, Politi K, et al. (2004) EGF receptor gene mutations are common in lung cancers from "never smokers" and are associated with

sensitivity of tumors to gefitinib and erlotinib. Proc Natl Acad Sci U S A 101: 13306-13311.

42. Bardelli A, Parsons DW, Silliman N, Ptak J, Szabo S, et al. (2003) Mutational analysis of the tyrosine kinome in colorectal cancers. Science 300: 949.

43. Rand V, Huang J, Stockwell T, Ferriera S, Buzko O, et al. (2005) Sequence survey of receptor tyrosine kinases reveals mutations in glioblastomas. Proc Natl Acad Sci U S A 102: 14344-14349.

44. Arjona D, Bello MJ, Alonso ME, Aminoso C, Isla A, et al. (2005) Molecular analysis of the EGFR gene in astrocytic gliomas: mRNA expression, quantitative-PCR analysis of non-homogeneous gene amplification and DNA sequence alterations. Neuropathol Appl Neurobiol 31: 384-394.

45. Kobayashi S, Ji H, Yuza Y, Meyerson M, Wong KK, et al. (2005) An alternative inhibitor overcomes resistance caused by a mutation of the epidermal growth factor receptor. Cancer Res 65: 7096-7101.

46. Sawyers CL (2005) Calculated resistance in cancer. Nat Med 11: 824-825.

47. Nishikawa R, Ji XD, Harmon RC, Lazar CS, Gill GN, et al. (1994) A mutant epidermal growth factor receptor common in human glioma confers enhanced tumorigenicity. Proc Natl Acad Sci U S A 91: 7727-7731.

48. Li S, Schmitz KR, Jeffrey PD, Wiltzius JJ, Kussie P, et al. (2005) Structural basis for inhibition of the epidermal growth factor receptor by cetuximab. Cancer Cell 7: 301-311.

ACKNOWLEDGEMENTS

We thank Dr. Elliot Landaw for statistical analysis, Dr. Neil Shah and Dr. Brian Skaggs for technical assistance. and Mr. Charlie Hatton, Mr. Rick Nicoletti, Ms. Meng Wang, and Ms. Megan Hanna for bioinformatics support. This work was supported by grants from Accelerate Brain Cancer Cure (Drs. Mellinghoff, Sawyers, Mischel), the Henry E. Singleton Brain Tumor Foundation (Drs. Mellinghoff, Mischel, Cloughesy), the Goldhirsch Foundation (Dr. Sawyers), the Phase One Foundation (Dr. Sawyers), the Brain Tumors Funders' Collaborative, the American Cancer Society (Dr. Meyerson), the National Cancer Institute (Drs. Meyerson and Sellers), the National Institutes of Health (R01 CA116020, Dr. Meyerson), the Novartis Foundation (Drs. Meyerson and Sellers), the Claudia Adams Barr program (Drs. Meyerson and Sellers), and the National Institute for Neurological Disorders and Stroke (Dr. Mischel). Ms. Huang is the recipient of a Medical Scientist Training Grant (David Geffen School of Medicine at UCLA). Dr. Thomas is a Mildred Scheel Fellow of the Deutsche Krebshilfe (German Cancer Aid). Dr. Beroukhim is the recipient of a Department of Defense postdoctoral training award. Dr. Sawyers is a Doris Duke Distinguished Clinical Investigator. Dr. Mellinghoff is a Forbeck Scholar and recipient of a Department of Defense Physician Scientist Training Award.

FIGURE LEGENDS

Figure 1. *EGFR* missense mutations in glioblastoma cluster in the extracellular domain and are associated with increased EGFR gene dose.

(A) Location of missense mutations within the EGF receptor protein in a panel of 151 gliomas (132 glioblastomas, eleven WHO grade III gliomas, and eight glioblastoma cell lines). Each diamond represents one sample harboring the indicated mutation. Amino acid (AA) numbers are based on the human EGFR precursor protein (accession number P00533). Ligand-binding domains (I and III), cysteine-rich domains (II and IV), kinase domain (kinase), and the extracellular deletion mutant *EGFR*vIII [47] are indicated as reference. (**B)** Increased *EGFR* gene dose in tumors harboring *EGFR* missense mutations. Left. High-resolution view of Affymetrix 100K SNP array at the *EGFR* gene locus for ten glioblastoma tumors and three normal controls (sample numbers are indicated above each column). *EGFR* mutation and estimated gene copy number are indicated below each column. Right. Comparison of EGFR gene copy number determination by SNP array (Y-axis, EGFR log 2 ratios) and FISH. AMP = amplified, NON-AMP = non amplified. **(C)** RT-PCR for EGFRvIII and full-length EGFR in fourteen fresh frozen glioblastoma tumors (see Methods). The upper band represents full-length EGFR (1044 bp), the lower band EGFRvIII (243 bp), and the inset shows glyceraldehyde-3-phosphate dehydrogenase (GAPDH) RT-PCR results.

Figure 2. EGFR missense mutations are transforming and tumorigenic

(A) Anchorage-independent growth of NIH-3T3 cells expressing various *EGFR* alleles. (mean number of colonies +/- standard deviation).The lower panel shows EGFR and actin immunoblots of whole cell lysates from NIH-3T3 subclones plated in soft agar. (**B)** Tumorigenicity of NIH-

Figure 1

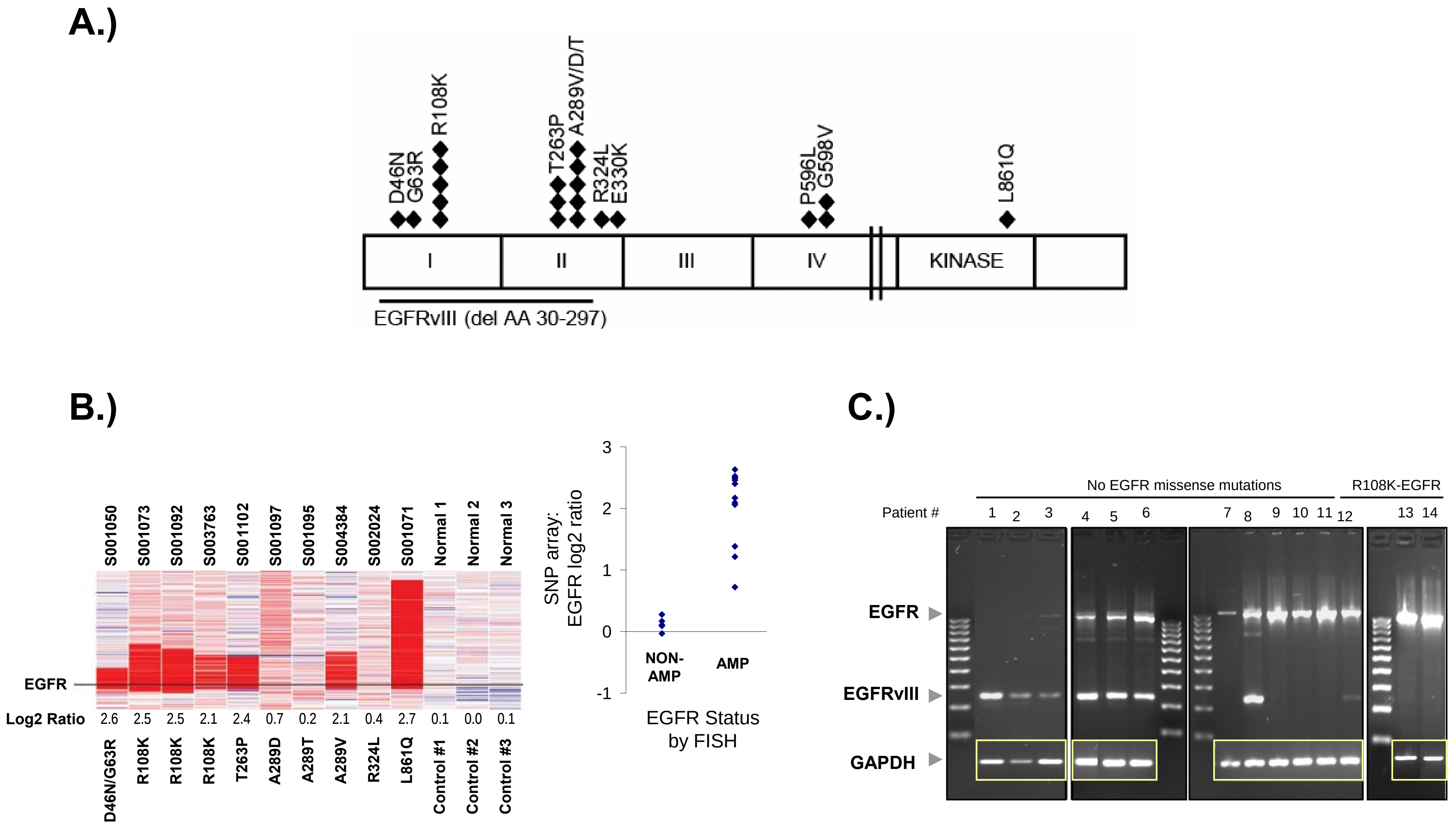

A.)
D46N
G63R
R108K
T263P
A289V/D/T
R324L
E330K
P596L
G598V
L861Q
I
II
III
IV
KINASE
EGFRvIII (del AA 30-297)
B.)
S001050
S001073
S001092
S003763
S001102
S001097
S001095
S004384
S002024
S001071
Normal 1
Normal 2
Normal 3
EGFR
Log2 Ratio
2.6
2.5
2.5
2.1
2.4
0.7
0.2
2.1
0.4
2.7
0.1
0.0
0.1
D46N/G63R
R108K
R108K
R108K
T263P
A289D
A289T
A289V
R324L
L861Q
Control #1
Control #2
Control #3
SNP array: EGFR log2 ratio
3
2
1
0
-1
NON-AMP
AMP
EGFR Status by FISH
C.)
No EGFR missense mutations
R108K-EGFR
Patient #
1
2
3
4
5
6
7
8
9
10
11
12
13
14
EGFR
EGFRvIII
GAPDH

Table 1

Sample #	Histology	Exon	Nucleotide Change	AA Change	Somatic	Mutation detection		Abundance of mutant allele[2]	EGFR gene dose		EGFRvIII
						Sanger	MALDI-TOF		FISH	SNP-array[3]	
S001050	GBM	2	G136A, G187C	D46N, G63R	Somatic	Het	+	81/95 (85.3 %)	AMP	2.63	negative
S001073	GBM	3	G323A	R108K	Somatic	Hom	+	88/90 (97.8 %)	AMP	2.52	negative
S001076	GBM	3	G323A	R108K	Somatic	Het[1]	+	10/82 (12.2 %)	NON AMP	n.d.	negative
S001092	GBM	3	G323A	R108K	Somatic	Het[1]	+	16/93 (17.2 %)	AMP	2.5	negative
S001094	GBM	3	G323A	R108K	Somatic	Het[1]	+	7/90 (7.8 %)	NON AMP	n.d.	negative
S003763	GBM	3	G323A	R108K	Unknown	Het[1]	+	3/94 (3.2 %)	AMP	2.058	positive
S001067	GBM	7	A787C	T263P	Somatic	Hom	+	83/92 (90.2 %)	n.d.	n.d.	n.d.
S001102	GBM	7	A787C	T263P	Somatic	Het	+	69/92 (75 %)	AMP	2.4	negative
S001103	GBM	7	A787C	T263P	Somatic	Het[1]	+	3/65 (4.6 %)	AMP	n.d.	negative
S001097	GBM	7	C866A	A289D	Unknown	Het	+	n.d.	AMP	0.72	negative
S001095	GBM	7	G865A	A289T	Unknown	Het	+	n.d.	NON AMP	0.17	negative
S001090	GBM	7	C866T	A289V	Somatic	Het[1]	+	3/82 (3.7 %)	NON AMP	n.d.	negative
S001108	GBM cell line	7	C866T	A289V	Unknown	Het	+	31/92 (33.7 %)	n.d.	n.d.	negative
S004384	GBM	7	C866T	A289V	Unknown	Het	+	83/90 (92.2 %)	AMP	2.1	negative
S002024	GBM	8	G971T	R324L	Unknown	Het	n.d.	n.d.	n.d.	0.4	negative
S001026	GBM	8	G988A	E330K	Germline	Het	+	38/94 (40.4 %)	NON AMP	n.d.	negative
S003577	GBM	15	C1787T	P596L	Somatic	Het	+	37/89 (41.6 %)	NON AMP	n.d.	negative
S001018	GBM	15	G1793T	G598V	Somatic	Het	n.d.	48/94 (51.1 %)	AMP	n.d.	negative
S001005	GBM	15	G1793T	G598V	Unknown	Hom	+	77/92 (83.7 %)	NON AMP	n.d.	n.d.
S001071	GBM	21	T2582A	L861Q	Somatic	Het	+	n.d.	AMP	2.49	n.d.

[1], detected with sub-threshold signal; [2], number of colonies with mutant/number of colonies with wildtype EGFR; [3], smoothened log2 ratio at the *EGFR* locus

Abbreviations: AA, amino acid; FISH, fluorescent in-situ hybridization; GBM, glioblastoma; Het, heterozygous; Hom, homozygous; MALDI-TOF, matrix-assisted laser desorption/ionization time-of-flight mass spectrometry; n.d., not determined (failed reaction or sample not available); SNP, single-nucleotide polymorphism

Figure 2

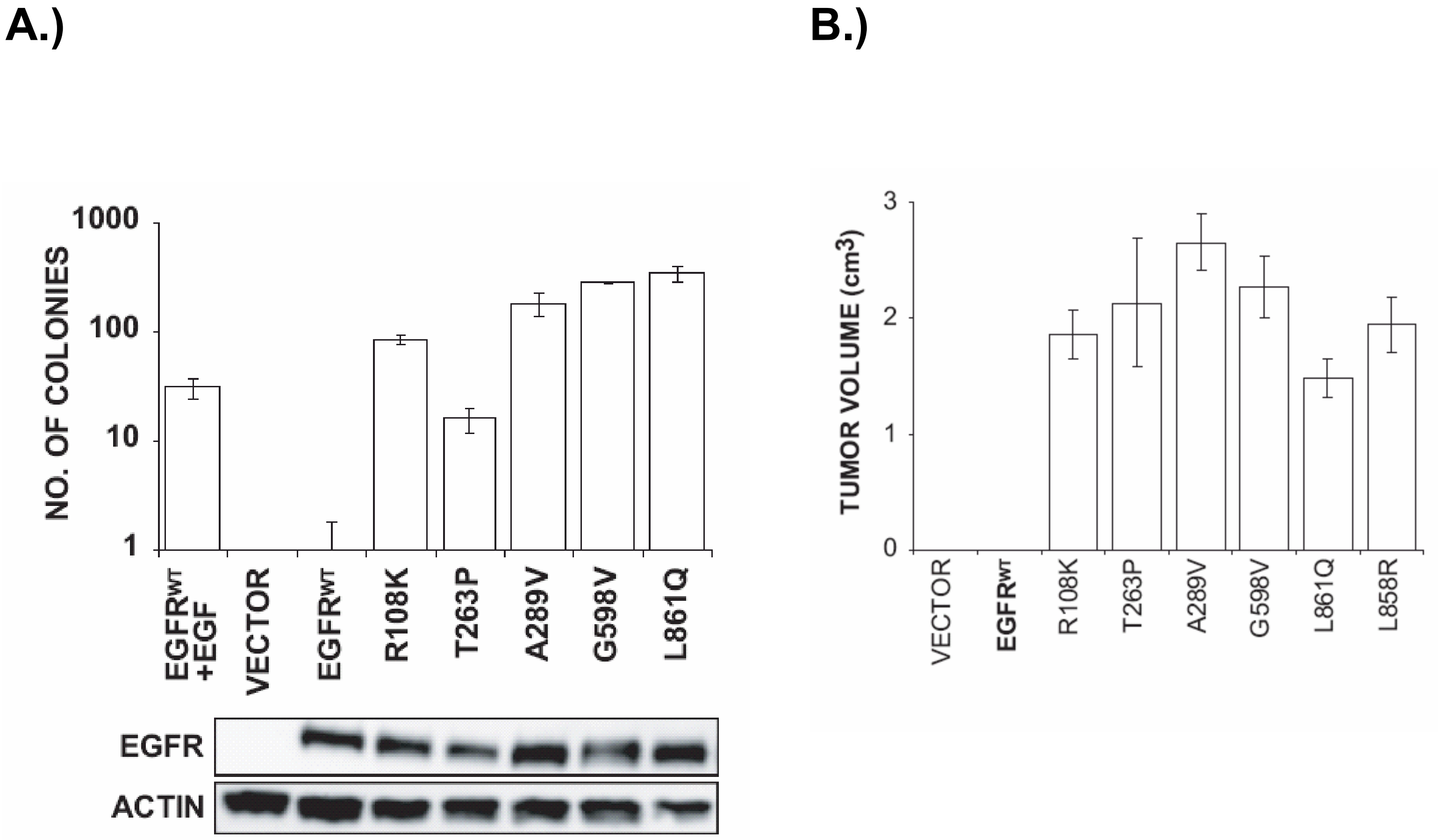

A.)
B.)
NO. OF COLONIES
1000
100
10
1
EGFR^{WT} +EGF
VECTOR
EGFR^{WT}
R108K
T263P
A289V
G598V
L861Q
EGFR
ACTIN
TUMOR VOLUME (cm^3)
3
2
1
0
VECTOR
EGFR^{WT}
R108K
T263P
A289V
G598V
L861Q
L858R

Figure 3

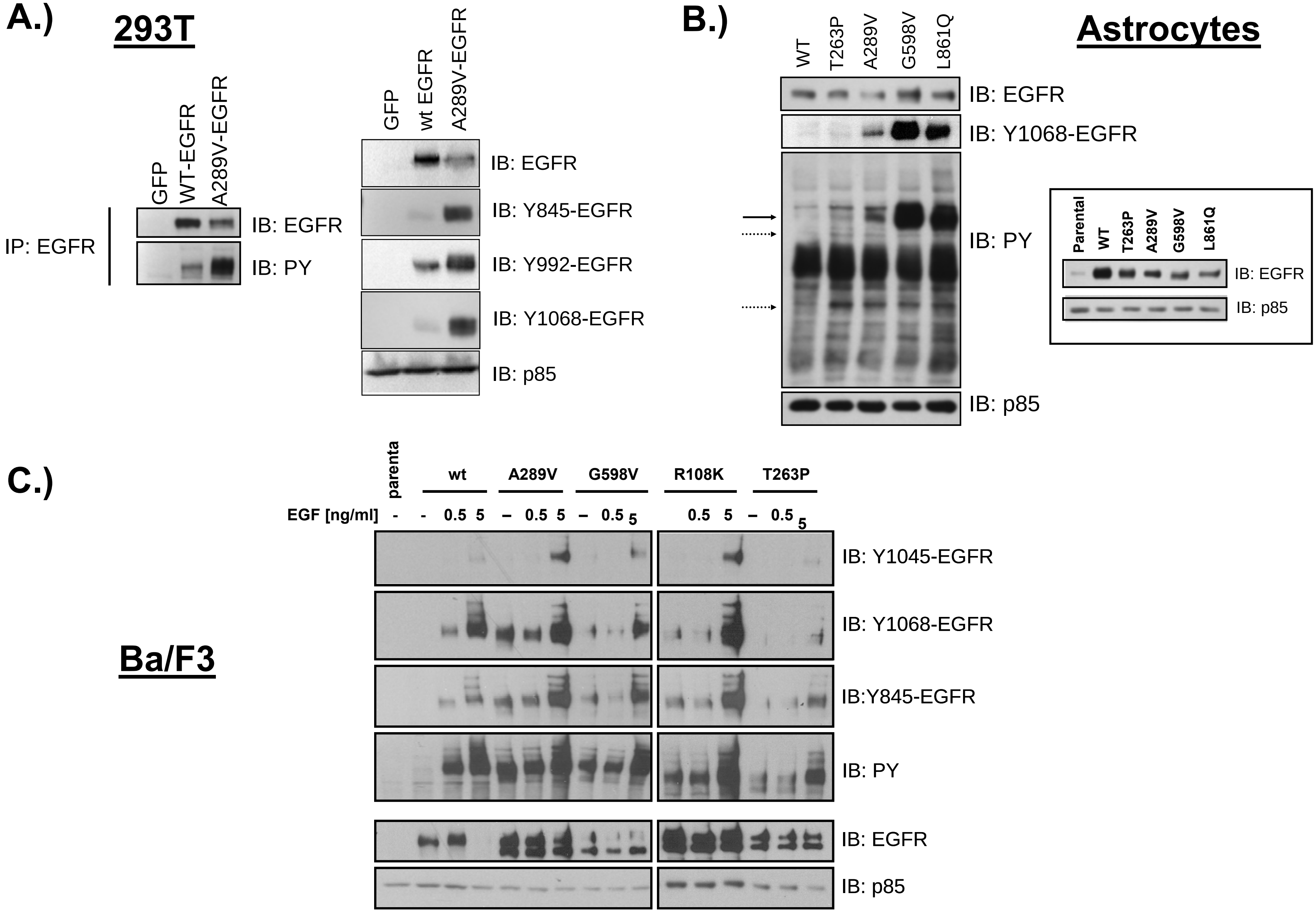
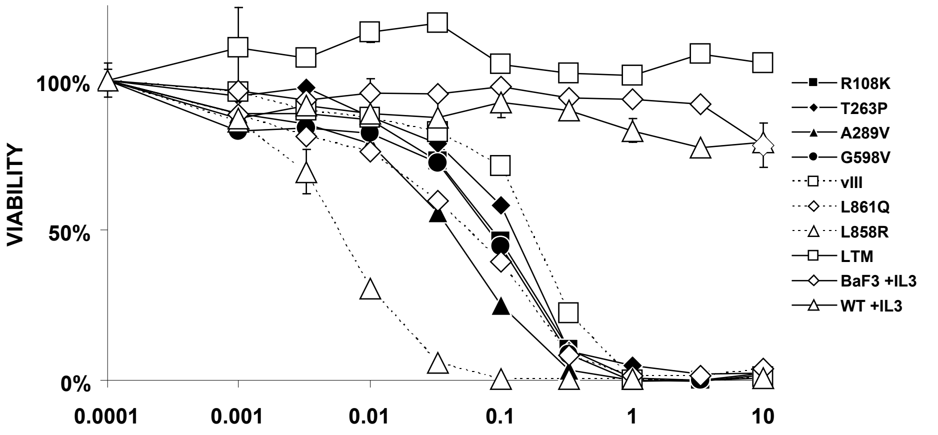
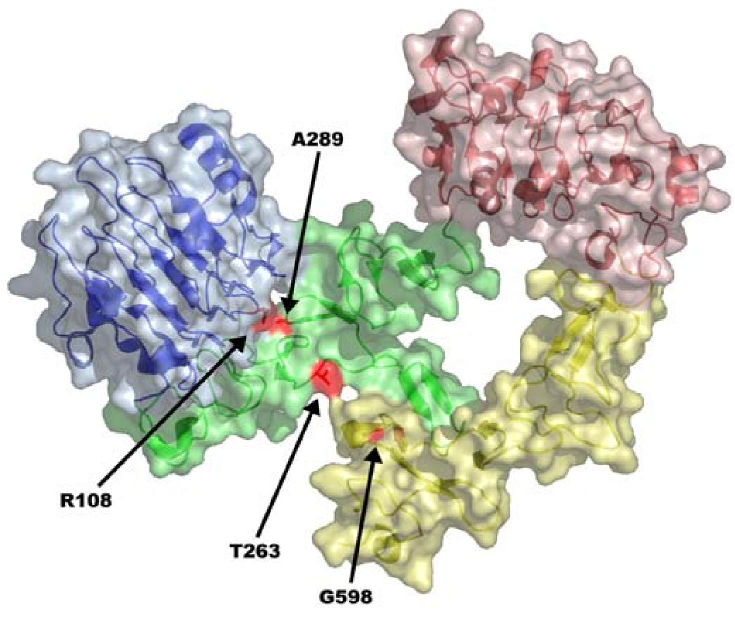

A.)
293T
GFP
WT-EGFR
A289V-EGFR
IP: EGFR
IB: EGFR
IB: PY
GFP
wt EGFR
A289V-EGFR
IB: EGFR
IB: Y845-EGFR
IB: Y992-EGFR
IB: Y1068-EGFR
IB: p85
B.)
WT
T263P
A289V
G598V
L861Q
Astrocytes
IB: EGFR
IB: Y1068-EGFR
IB: PY
IB: p85
Parental
WT
T263P
A289V
G598V
L861Q
IB: EGFR
IB: p85
C.)
parenta
wt
A289V
G598V
R108K
T263P
EGF [ng/ml]
-
-
0.5
5
–
0.5
5
–
0.5
5
0.5
5
–
0.5
5
Ba/F3
IB: Y1045-EGFR
IB: Y1068-EGFR
IB:Y845-EGFR
IB: PY
IB: EGFR
IB: p85

Figure 4

A.)


VIABILITY
100%
50%
0%
0.0001
0.001
0.01
0.1
1
10
R108K
T263P
A289V
G598V
vIII
L861Q
L858R
LTM
BaF3 +IL3
WT +IL3

B.)


A289
R108
T263
G598

Supplementary Figure S1

D46N G63R

```
HOMO SAPIENS EGFR             KVCQGTSNKLTQLGTFEDHFLSLQRMFNNCEVVLGN-LEITYVQ-RNYDL 76
HOMO SAPIENS HER2 (ERBB2)     QVCTGTDMKLRLPASPETHLDMLRHLYQGCQVVQGN-LELTYLP-TNASL 71
HOMO SAPIENS HER3 (ERBB3)     AVCPGTLNGLSVTGDAENQYQTLYKLYERCEVVMGN-LEIVLTG-HNADL 74
HOMO SAPIENS ERBB4            SVCAGTENKLSSLSDLEQQYRALRKYYENCEVVMGN-LEITSIE-HNRDL 74
MUS MUSCULUS                  KVCQGTSNRLTQLGTFEDHFLSLQRMYNNCEVVLGN-LEITYVQ-RNYDL 76
RATTUS NORVEGICUS             KVCQGTSNRLTQLGTFEDHFLSLQRMFNNCEVVLGN-LEITYVQ-RNYDL 76
SUS SCROFA                    KVCQGTSNKLTQLGTFEDHFLSLQRMFNNCEVVLGN-LEITYMQ-NSYNL 76
DANIO RERIO                   KVCQGANNKLTLLGTVEDHYQVLLRMYRNCTVVLEN-LEITHIT-EKYDL 73
DROSOPHILA MELANOGASTER       KVCIGTKSRLSVPSNKEHHYRNLRDRYTNCTYVDGN-LELTWLPNENLDL 99
XIPHOPHORUS (XMRK)            KVCQGTSNQMTML---DNHYLKMKKMYSGCNVVLENTLEITYTQ-ENQDL 73
                               ** *:   :      : :   :   :  *  *  * **:.     . .*
```

R108K

```
HOMO SAPIENS EGFR             SFLKTIQEVAGYVLIALNTVERIPLENLQIIRGNMYYENSYALAVLSNYD 126
HOMO SAPIENS HER2 (ERBB2)     SFLQDIQEVQGYVLIAHNQVRQVPLQRLRIVRGTQLFEDNYALAVLDNGD 121
HOMO SAPIENS HER3 (ERBB3)     SFLQWIREVTGYVLVAMNEFSTLPLPNLRVVRGTQVYDGKFAIFVMLNYN 124
HOMO SAPIENS ERBB4            SFLRSVREVTGYVLVALNQFRYLPLENLRIIRGTKLYEDRYALAIFLNYR 124
MUS MUSCULUS                  SFLKTIQEVAGYVLIALNTVERIPLENLQIIRGNALYENTYALAILSNYG 126
RATTUS NORVEGICUS             SFLKTIQEVAGYVLIALNTVERIPLENLQIIRGNALYENTYALAVLSNYG 126
SUS SCROFA                    SFLKTIQEVAGYVLIALNTVEKIPLENLQIIRGNVLYENTHALAVLSNYG 126
DANIO RERIO                   SFLKSIQEVGGYVLIAVNTVSKIPLENLRIIRGHSLYEDKFALAVLVNFN 123
DROSOPHILA MELANOGASTER       SFLDNIREVTGYILISHVDVKKVVFPKLQIIRGRTLFSLSVEEEKYALFV 149
XIPHOPHORUS (XMRK)            SFLQSIQEVGGYVLIAMNEVSTIPLVNLRLIRGQNLYEGNFTLLVMSNYQ 123
                              ***  ::** **:*::   .  : : .*:::**   :.
```

T263P

```
HOMO SAPIENS EGFR             QQCSG-RCRGKSPSDCCHNQCAAGCTGPRESDCLVCRKFRDEATCKDTCP 265
HOMO SAPIENS HER2 (ERBB2)     GGCA--RCKGPLPTDCCHEQCAAGCTGPKHSDCLACLHFNHSGICELHCP 269
HOMO SAPIENS HER3 (ERBB3)     PQCNG-HCFGPNPNQCCHDECAGGCSGPQDTDCFACRHFNDSGACVPRCP 260
HOMO SAPIENS ERBB4            EQCDG-RCYGPYVSDCCHRECAGGCSGPKDTDCFACMNFNDSGACVTQCP 263
MUS MUSCULUS                  QQCSH-RCRGRSPSDCCHNQCAAGCTGPRESDCLVCQKFQDEATCKDTCP 265
RATTUS NORVEGICUS             QQCSR-RCRGRSPSDCCHNQCAAGCTGPRESDCLVCHRFRDEATCKDTCP 265
SUS SCROFA                    QQCSG-RCRGRSPSDCCHNQCAAGCTGPRESDCLVCRRFRDEATCKDTCP 265
DANIO RERIO                   EQCSG-RCKGPRPIDCCNEHCAAGCTGPRPTDCLACKDFQDEGTCKDACP 263
DROSOPHILA MELANOGASTER       PQCAGGRCYGPKPRECCHLFCAGGCTGPTQKDCIACKNFFDEGVCKEECP 287
XIPHOPHORUS (XMRK)            EQCNR-RCRGPKPIDCCNEHCAGGCTGPRATDCLACRDFNDDGTCKDTCP 268
                                *   :* *    :**:  **.**:**  .**:.*  * ... *   **
```

A289D/T/V

```
HOMO SAPIENS EGFR             PIMLYNPTTYQMDVNPEGKYSFGATCVKKCPRNYVVTDHGSCVRACGADS 315
HOMO SAPIENS HER2 (ERBB2)     ALVTYNTDTFESMPNPEGRYTFGASCVTACPYNYLSTDVGSCTLVCPLHN 319
HOMO SAPIENS HER3 (ERBB3)     QPLVYNKLTFQLEPNPHTKYQYGGVCVASCPHNFVVDQT-SCVRACPPDK 309
HOMO SAPIENS ERBB4            QTFVYNPTTFQLEHNFNAKYTYGAFCVKKCPHNFVVDSS-SCVRACPSSK 312
MUS MUSCULUS                  PIMLYNPTTYQMDVNPEGKYSFGATCVKKCPRNYVVTDHGSCVRACGPDY 315
RATTUS NORVEGICUS             PIMLYNPTTYQMDVNPEGKYSFGATCVKKCPRNYVVTDHGSCVRACGPDY 315
SUS SCROFA                    PIMLYNPTTYQMDVNPLGKYSFGATCVKKCPRNYVVTDHGSCVRACSSDS 315
DANIO RERIO                   RIMLYDPNTHQLAPNPYGKYSFGATCIKTCPHNYVVTDHGACVRTCSPGT 313
DROSOPHILA MELANOGASTER       PMRKYNPTTYVLETNPEGKYAYGATCVKECPG-HLLRDNGACVRSCPQDK 336
XIPHOPHORUS (XMRK)            PPKIYDIVSHQVVDNPNIKYTFGAACVKECPSNYVVTEG-ACVRSCSAGM 317
                                  *:  :.    *   :* :*. *:  **  .:  .  :*.  *
```

R324L E330K

```
HOMO SAPIENS EGFR             YEME-EDGVRKCKKCEGPCRKVCNGIGIGEFKDSLSINATNIKHFKNCTS 364
HOMO SAPIENS HER2 (ERBB2)     QEVTAEDGTQRCEKCSKPCARVCYGLGMEHLREVRAVTSANIQEFAGCKK 369
HOMO SAPIENS HER3 (ERBB3)     MEVD-KNGLKMCEPCGGLCPKACEGTGSGSRF--QTVDSSNIDGFVNCTK 356
HOMO SAPIENS ERBB4            MEVE-ENGIKMCKPCTDICPKACDGIGTGSLMSAQTVDSSNIDKFINCTK 361
MUS MUSCULUS                  YEVE-EDGIRKCKKCDGPCRKVCNGIGIGEFKDTLSINATNIKHFKYCTA 364
RATTUS NORVEGICUS             YEVE-EDGVSKCKKCDGPCRKVCNGIGIGEFKDTLSINATNIKHFKYCTA 364
SUS SCROFA                    YEVE-EDGVRKCKKCDGPCGKVCNGIGIGEFKDTLSINATNIKHFRNCTS 364
DANIO RERIO                   YEVD-EGGVRKCKRCEGLCPKVCNGLGMGPLANVLSINATNIDSFENCTK 362
DROSOPHILA MELANOGASTER       MDKG-----GECVPCNGPCPKTCPGVT--------VLHAGNIDSFRNCTV 373
XIPHOPHORUS (XMRK)            LEVD-ENGKRSCKPCDGVCPKVCDGIGIGSLSNTIAVNSTNIGSFSNCTK 366
                               :         *  *   * :.* *           : : **  *  *.
```

P596L

```
HOMO SAPIENS EGFR             CIQCHPECLPQAM-NITCTGRGPDNCIQCAHYIDGPHCVKTCP------- 596
HOMO SAPIENS HER2 (ERBB2)     CLPCHPECQPQNG-SVTCFGPEADQCVACAHYKDPPFCVARCP------- 601
HOMO SAPIENS HER3 (ERBB3)     CFSCHPECQPMEG-TATCNGSGSDTCAQCAHFRDGPHCVSSCP------- 590
HOMO SAPIENS ERBB4            CVECDPQCEKMEDGLLTCHGPGPDNCTKCSHFKDGPNCVEKCP------- 594
MUS MUSCULUS                  CIQCHPECLPQAM-NITCTGRGPDNCIQCAHYIDGPHCVKTCP------- 596
RATTUS NORVEGICUS             CIQCHPECLPQTM-NITCTGRGPDNCIKCAHYVDGPHCVKTCP------- 596
SUS SCROFA                    CVQCHPECLPQAK-NVTCMGRGPDSCVRCAHYIDGPHCVKTCP------- 596
DANIO RERIO                   CMECDPECLIMNE-TQTCNGPGPDKCTVCANYKDGPHCVHRCP------- 594
DROSOPHILA MELANOGASTER       CKICHPECR-------TCNGAGADHCQECVHVRDGQHCVSECPKNKYNDR 613
XIPHOPHORUS (XMRK)            CVQCHQECLVQTD-SLTCYGPGPANCSKCAHFQDGPQCIPRCP------- 591
                              *  *. :*        ** *  .  *  * :  *   *:  **
```

G598V

```
HOMO SAPIENS EGFR             ----------AGVMGENNTL-V---------------------------- 607
HOMO SAPIENS HER2 (ERBB2)     ----------SGVKPDLSYMPI---------------------------- 613
HOMO SAPIENS HER3 (ERBB3)     ----------HGVLGAKGPI------------------------------ 600
HOMO SAPIENS ERBB4            ----------DGLQGANSFI------------------------------ 604
MUS MUSCULUS                  ----------AGIMGENNTL-V---------------------------- 607
RATTUS NORVEGICUS             ----------SGIMGENNTL-V---------------------------- 607
SUS SCROFA                    ----------AGIAGENSTL-I---------------------------- 607
DANIO RERIO                   ----------QGVPGEKDTL-I---------------------------- 605
DROSOPHILA MELANOGASTER       GVCRECHATCDGCTGPKDTIGIGACTTCNLAIINNDATVKRCLLKDDKCP 663
XIPHOPHORUS (XMRK)            ----------HGMLGDGDTL-I---------------------------- 602
                                         *     . :
```

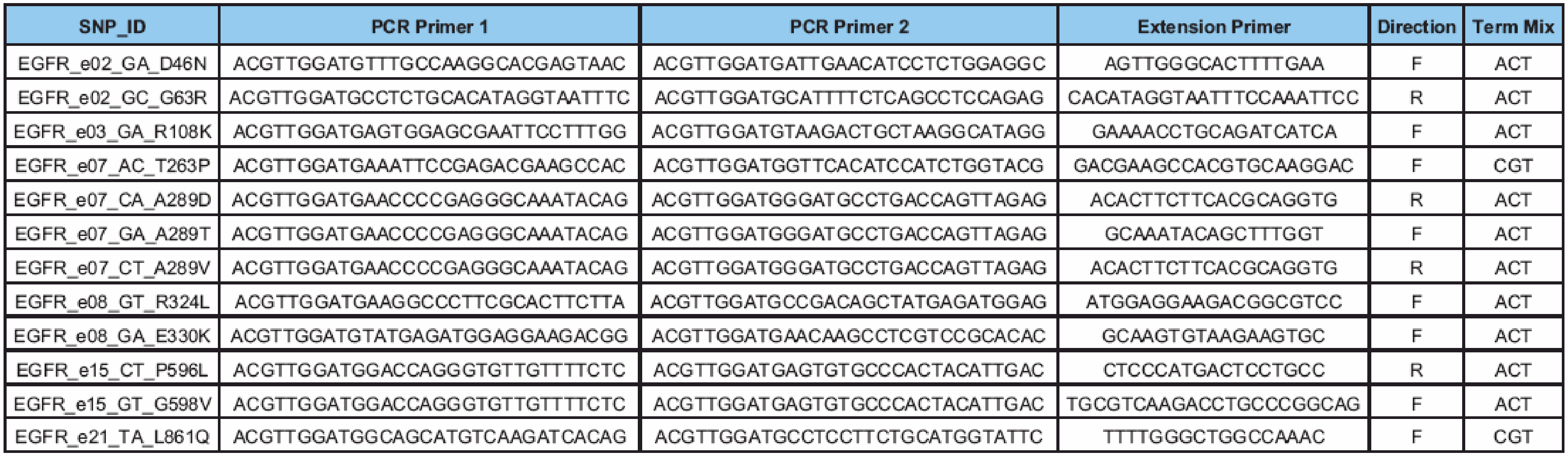
L861Q

```
HOMO SAPIENS EGFR             RDLAARNVLVKTPQHVKITDFGLAKLLGAEEKEYHAEGGKVPIKWMALES 885
HOMO SAPIENS HER2 (ERBB2)     RDLAARNVLVKSPNHVKITDFGLARLLDIDETEYHADGGKVPIKWMALES 893
HOMO SAPIENS HER3 (ERBB3)     RNLAARNVLLKSPSQVQVADFGVADLLPPDDKQLLYSEAKTPIKWMALES 882
HOMO SAPIENS ERBB4            RDLAARNVLVKSPNHVKITDFGLARLLEGDEKEYNADGGKMPIKWMALEC 891
MUS MUSCULUS                  RDLAARNVLVKTPQHVKITDFGLAKLLGAEEKEYHAEGGKVPIKWMALES 887
RATTUS NORVEGICUS             RDLAARNVLVKTPQHVKITDFGLAKLLGAEEKEYHAEGGKVPIKWMALES 886
SUS SCROFA                    RDLAARNVLVKTPQHVKITDFGLAKLLGAEEKEYHAEGGKVPIKWLALES 885
DANIO RERIO                   RDLAARNVLVKTPQHVKITDFGLAKLLNADEKEYHADGGKVPIKWMALES 884
DROSOPHILA MELANOGASTER       RDLAARNVLVQTPSLVKITDFGLAKLLSSDSNEYKAAGGKMPIKWLALEC 1062
XIPHOPHORUS (XMRK)            RDLAARNVLLKNPNHVKITDFGLSKLLTADEKEYQADGGKVPIKWMALES 880
                              *:*******::.*. *:::***:: **  :..:     .* ****:***.
```

Supplementary Table S1

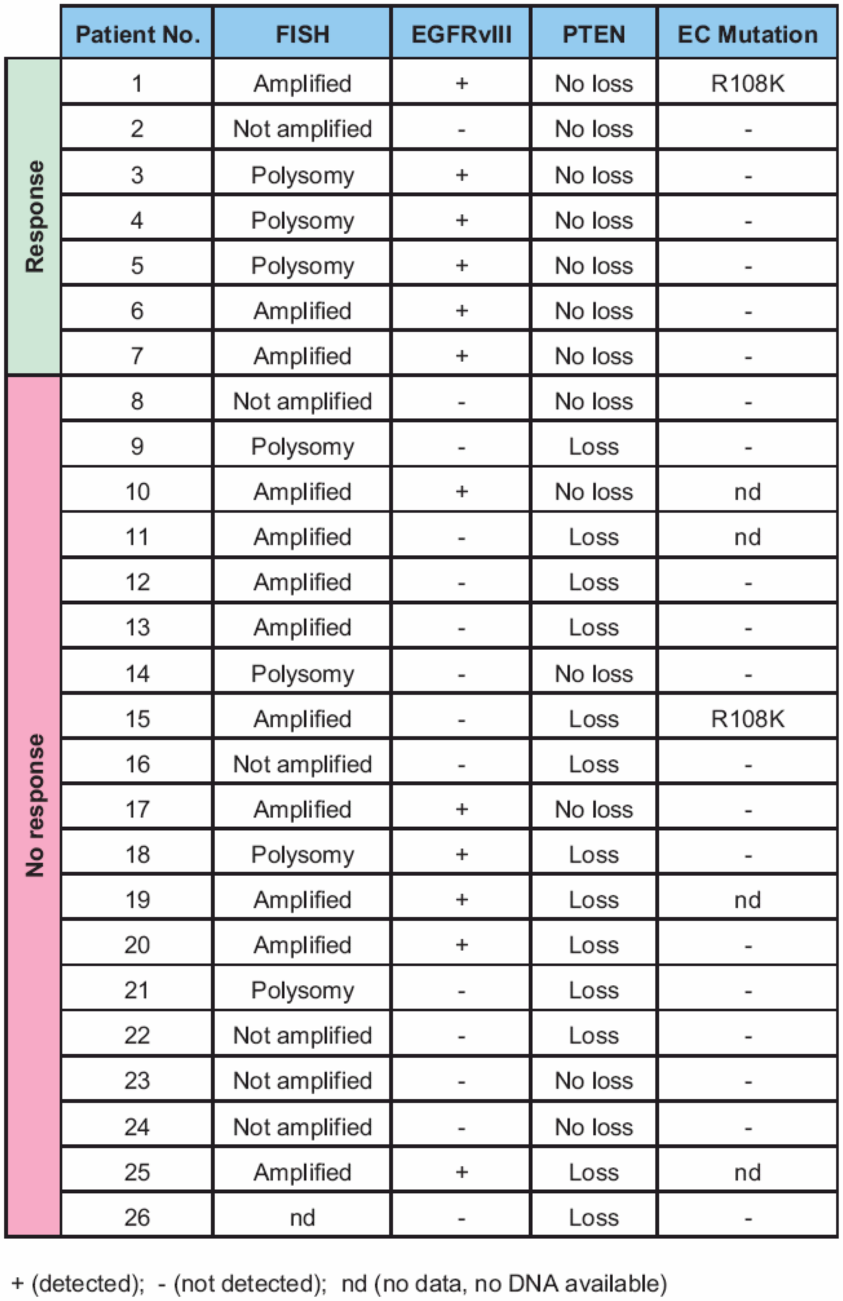
SNP_ID	PCR primer 1	PCR primer 2	Extension Primer	Direction	Term Mix
rs464221	ACGTTGGATGTGAGGACTTGGGATTAGGAC	ACGTTGGATGAGGCAGCAGCAGAAGTTTAG	CAGTTAAGAATCATCCAACC	F	ACT
rs1385306	ACGTTGGATGGGCTCTGTATAGAGCTTTGG	ACGTTGGATGTCTGTCAGCTGTTCTGATGC	TTGGCTCATCTGTTCTC	R	ACT
rs1434199	ACGTTGGATGGGCATAGGGAGCTGAATCAA	ACGTTGGATGCACCTCTGCCCCCTAATTTC	GAGCTGAATCAATACATATCACT	F	ACT
rs1259859	ACGTTGGATGTAGTATCAAGGTTTTCTGGC	ACGTTGGATGTGCCAATTTTATAGAAACTC	TGGAGGTGATCCTTCTTATA	R	ACT
rs120434	ACGTTGGATGATGCCTTGTTCCATGTGCTG	ACGTTGGATGACACAAGTGGAAGCTTGCAG	TGTGCTGGAATGCTGAT	R	ACT
rs1365740	ACGTTGGATGACTTCCTATGTTGCCAGCAC	ACGTTGGATGGTACTTACCTCCTGAGGTAG	CACCTTTTTCCTCTTCATT	F	ACT
rs265005	ACGTTGGATGACACAAAGCTACCACTGCAC	ACGTTGGATGATCAGGCACAATCTCTACCC	GTCTGCTCAACTGGACTATAAA	R	ACT
rs248205	ACGTTGGATGGTACTTAGCTCTATTGTGGG	ACGTTGGATGGGCTCATGGGAAGAATCATC	AAAAGCTCCAACACACT	F	ACT
rs1350836	ACGTTGGATGATTTCAGACTGTTGTGCTGG	ACGTTGGATGCCTGTGGCCTTTTCATGGAG	TGTTGTGCTGGGTCTTG	R	ACT
rs2347790	ACGTTGGATGACCTATGCAACAGCTGGAAG	ACGTTGGATGATGCTCCACCCACACAGTTC	GAAGATGTGCGTATGCCTTA	F	ACT
rs2180770	ACGTTGGATGAACCACAGTGAGCATGGAAG	ACGTTGGATGCACTTGTCACAAACTCTAGG	AAATATAACCTTGATCCTCTG	R	ACT
rs1522307	ACGTTGGATGTAGGGTCCAGAAATGTGTTG	ACGTTGGATGGGAGAAGCAAGCCATAGATG	GCACAGCTTAAAGGTCTC	R	ACT
rs1390272	ACGTTGGATGCAGGATAGTCTACTATGTGC	ACGTTGGATGTCCTTTGATATTTGTTCCAC	GGTTTTTCTTAACAAGTTCAC	R	ACT
rs2016207	ACGTTGGATGCTGATGCAAAAATCTATGGC	ACGTTGGATGGCTCTAGTAGATTGCTAGCC	GCAAAAATCTATGGCTTTCATCAT	R	ACT
rs1549944	ACGTTGGATGCTCTCTTGAAGAAGGTGCAG	ACGTTGGATGCACATATGCTGGCCTTGTTC	GGAAAGAAACACTGATCCA	F	ACT
AMELXY	ACGTTGGATGTGAGCTGGCACCACTGGGAT	ACGTTGGATGAAATCATCCCCGTGCTGTCC	CTGGGATGTGGTGATGAG	R	ACT
rs1950501	ACGTTGGATGAATGACTCACCACTGACCAC	ACGTTGGATGTGTTGTCTGGGAACTCAGGG	GACCACTGCTTTTTATGC	F	ACT
rs2051068	ACGTTGGATGTCTGAAGGTAAAGATTCAAG	ACGTTGGATGCTGGAAACCATTTCAGAATGC	AGAGCGCACAGGTATAG	R	ACT
rs1028330	ACGTTGGATGCTCTACATCTGTGGGTCTAG	ACGTTGGATGTTTTCGGGCAGTGAAGAGAC	GACATGTATCCACCATTATG	R	ACT
rs2380657	ACGTTGGATGGCTAGGGTTGAAAACCAATG	ACGTTGGATGGAATATGAGCACAACACACG	CAGAAATAATGAGTGGAGA	R	ACT
rs1952966	ACGTTGGATGGGGCTAAGTGGAGAAGTTTG	ACGTTGGATGGCTGAAGTTTCACTTTATCC	CTTGGCTGAAGCAATAC	F	ACT
rs298898	ACGTTGGATGCCATATATACAGGGTGTATAG	ACGTTGGATGGTGTAGTGAGTGCTATGATC	TAACAAAGGAAGAAATGGGATA	R	ACT
rs1571256	ACGTTGGATGGCATTCTAAACATGCCTCTC	ACGTTGGATGATACCGCAGAAGTTTGGCAC	CCTCTCAGTACTCTAGTC	R	ACT
rs58616	ACGTTGGATGTAGGCAGAAAAGGGCTGAAG	ACGTTGGATGCCCTTCTGCTTTAACACTATC	GATGCTTTGTTTCAAGGTTA	F	ACT

Supplementary Table S2



SNP_ID	PCR Primer 1	PCR Primer 2	Extension Primer	Direction	Term Mix
EGFR_e02_GA_D46N	ACGTTGGATGTTTGCCAAGGCACGAGTAAC	ACGTTGGATGATTGAACATCCTCTGGAGGC	AGTTGGGCACTTTTGAA	F	ACT
EGFR_e02_GC_G63R	ACGTTGGATGCCTCTGCACATAGGTAATTTC	ACGTTGGATGCATTTTCTCAGCCTCCAGAG	CACATAGGTAATTTCCAAATTCC	R	ACT
EGFR_e03_GA_R108K	ACGTTGGATGAGTGGAGCGAATTCCTTTGG	ACGTTGGATGTAAGACTGCTAAGGCATAGG	GAAAACCTGCAGATCATCA	F	ACT
EGFR_e07_AC_T263P	ACGTTGGATGAAATTCCGAGACGAAGCCAC	ACGTTGGATGGTTCACATCCATCTGGTACG	GACGAAGCCACGTGCAAGGAC	F	CGT
EGFR_e07_CA_A289D	ACGTTGGATGAACCCCGAGGGCAAATACAG	ACGTTGGATGGGATGCCTGACCAGTTAGAG	ACACTTCTTCACGCAGGTG	R	ACT
EGFR_e07_GA_A289T	ACGTTGGATGAACCCCGAGGGCAAATACAG	ACGTTGGATGGGATGCCTGACCAGTTAGAG	GCAAATACAGCTTTGGT	F	ACT
EGFR_e07_CT_A289V	ACGTTGGATGAACCCCGAGGGCAAATACAG	ACGTTGGATGGGATGCCTGACCAGTTAGAG	ACACTTCTTCACGCAGGTG	R	ACT
EGFR_e08_GT_R324L	ACGTTGGATGAAGGCCCTTCGCACTTCTTA	ACGTTGGATGCCGACAGCTATGAGATGGAG	ATGGAGGAAGACGGCGTCC	F	ACT
EGFR_e08_GA_E330K	ACGTTGGATGTATGAGATGGAGGAAGACGG	ACGTTGGATGAACAAGCCTCGTCCGCACAC	GCAAGTGTAAGAAGTGC	F	ACT
EGFR_e15_CT_P596L	ACGTTGGATGGACCAGGGTGTTGTTTTCTC	ACGTTGGATGAGTGTGCCCACTACATTGAC	CTCCCATGACTCCTGCC	R	ACT
EGFR_e15_GT_G598V	ACGTTGGATGGACCAGGGTGTTGTTTTCTC	ACGTTGGATGAGTGTGCCCACTACATTGAC	TGCGTCAAGACCTGCCCGGCAG	F	ACT
EGFR_e21_TA_L861Q	ACGTTGGATGGCAGCATGTCAAGATCACAG	ACGTTGGATGCCTCCTTCTGCATGGTATTC	TTTTGGGCTGGCCAAAC	F	CGT

Supplementary Table S3



	Patient No.	FISH	EGFRvIII	PTEN	EC Mutation
Response	1	Amplified	+	No loss	R108K
	2	Not amplified	-	No loss	-
	3	Polysomy	+	No loss	-
	4	Polysomy	+	No loss	-
	5	Polysomy	+	No loss	-
	6	Amplified	+	No loss	-
	7	Amplified	+	No loss	-
No response	8	Not amplified	-	No loss	-
	9	Polysomy	-	Loss	-
	10	Amplified	+	No loss	nd
	11	Amplified	-	Loss	nd
	12	Amplified	-	Loss	-
	13	Amplified	-	Loss	-
	14	Polysomy	-	No loss	-
	15	Amplified	-	Loss	R108K
	16	Not amplified	-	Loss	-
	17	Amplified	+	No loss	-
	18	Polysomy	+	Loss	-
	19	Amplified	+	Loss	nd
	20	Amplified	+	Loss	-
	21	Polysomy	-	Loss	-
	22	Not amplified	-	Loss	-
	23	Not amplified	-	No loss	-
	24	Not amplified	-	No loss	-
	25	Amplified	+	Loss	nd
	26	nd	-	Loss	-

+ (detected); - (not detected); nd (no data, no DNA available)

Single nucleotide polymorphism array analysis of cancer

Amit Dutt[a,b] and Rameen Beroukhim[a,b,c]

Purpose of review

Classifying tumors and identifying therapeutic targets requires a description of the genetic changes underlying cancer. Single nucleotide polymorphism (SNP) arrays provide a high-resolution platform for describing several types of genetic changes simultaneously. With the resolution of these arrays increasing exponentially, they are becoming increasingly powerful tools for describing the genetic events underlying cancer.

Recent findings

The ability to map loss of heterozygosity (LOH) and overall copy number variations using SNP arrays is known. Techniques have recently been developed to map LOH at high resolution in the absence of paired normal data. Copy number variations described by SNP array studies are now reaching resolutions enabling the identification of novel oncogenes and tumor suppressor genes. The ability to determine allele-specific copy number changes has only recently been described. Moreover, SNP arrays offer a high-throughput platform for large-scale association studies that are likely to lead to the identification of multiple germline variants that predispose to cancer.

Summary

SNP arrays are an ideal platform for identifying both somatic and germline genetic variants that lead to cancer. They provide a basis for DNA-based cancer classification and help to define the genes being modulated, improving understanding of cancer genesis and potential therapy targets.

Keywords

cancer, copy number variation, loss of heterozygosity, SNP array

Curr Opin Oncol 18:000–000. © 2006 Lippincott Williams & Wilkins.

[a]Broad Institute of Harvard and Massachusetts Institute of Technology, Cambridge, Massachusetts, USA, [b]Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, Massachusetts, USA and [c]Departments of Medicine, Brigham and Women's Hospital, Boston, Massachusetts, USA

Correspondence to Rameen Beroukhim, Broad Institute of Harvard and Massachusetts Institute of Technology, 7CC, Cambridge, MA 02142, USA
Tel: +1 617 324 1582; e-mail: rameen@broad.mit.edu

Current Opinion in Oncology 2006, 18:000–000

Abbreviations

FFPE	formalin-fixed paraffin-embedded
LOH	loss of heterozygosity
PCR	polymerase chain reaction
SNP	single nucleotide polymorphism
TSG	tumor suppressor gene

1040-8746

Introduction

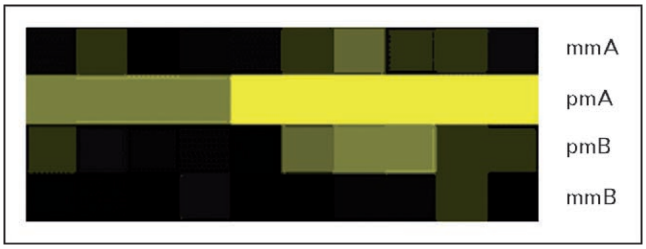
The understanding that cancer is a set of genetic diseases has led to increasing efforts to describe the genetic alterations underlying different cancers, both to identify therapeutic targets and to develop a robust classification system that reflects underlying biology. The somatic genetic alterations underlying cancer divide into several types, including point mutations and small insertion/deletion events, translocations, copy number changes, and loss of heterozygosity (LOH). Single nucleotide polymorphism (SNP) arrays offer the ability to define simultaneously the copy number changes and LOH events occurring in a tumor, at high resolution and throughout the genome, according to the two papers originally describing copy number measurements from SNP array [1,2]. As such, they offer a powerful and increasingly popular platform for oncogene and tumor suppressor gene (TSG) discovery, as well as cancer classification. In addition, a new generation of germline association studies using SNP arrays is likely to identify risk factors and biological mechanisms for the development of cancer. In this review, we will describe currently available high-density SNP array technologies, loss of heterozygosity analysis methods and applications, copy number analysis including allele-specific copy numbers, and the potential use of these arrays in identifying cancer-predisposing germline variants.

Background to SNP arrays

Accumulations of point mutations during evolution, in concert with random selection, have made SNPs the most common form of genetic variation in the human genome. A single base polymorphism is referred to as an SNP when the frequency of the minor allele exceeds 1% in at least one population; otherwise it is considered a variant or mutation [3]. It is estimated that there exist about 10 million SNPs throughout the genome, for an average of one SNP every 400–1000 base pairs [3]. Currently, about 5.6 million have been typed (dbSNP Build ID: 126), about half of which are estimated to have a minor allele frequency over 10% [4].

SNPs on a small chromosomal segment tend to be transmitted as a block, forming a haplotype. This correlation between alleles at nearby sites is known as linkage disequilibrium [5], and enables the prediction of the genotypes at a large number of SNP loci from known genotypes at a smaller number of representative SNPs, called 'tag SNPs' or 'haplotype tag SNPs' [6]. This reduction in the complexity of genetic variation between

Figure 1 View of a probe set for a single nucleotide polymorphism showing a homozygous 'A' call



On this array, 40 oligonucleotide probes are tiled for each SNP being interrogated, including perfect match (pm) and mismatch (mm) probes for each allele. The SNP position slides 5′ to 3′ among the probes. The fluorescence pattern indicates which alleles are present; the intensity indicates the quantity of bound DNA.

individuals enables much more efficient and economical determination of an individual's overall genotype: roughly 500 000 tag SNPs are adequate to fully genotype an individual with European ancestry [7].

For the study of germline genetic susceptibility to complex diseases, oligonucleotide arrays have been developed to interrogate such large numbers of SNP markers in a high-throughput, highly parallel fashion [8,9••,10••]. These arrays specifically detect the two different alleles of each SNP (an example is shown in Fig. 1). The use of these arrays in mapping somatic alterations in cancer, as opposed to germline variations in normal tissue, was suggested by three features: (i) their genotyping ability allows for analysis of LOH; (ii) for some of these arrays, copy number variations can be determined from signal intensities reflecting levels of DNA hybridization; and (iii) the density of SNP loci being interrogated allows for very high-resolution analysis.

High-throughput genotyping technologies

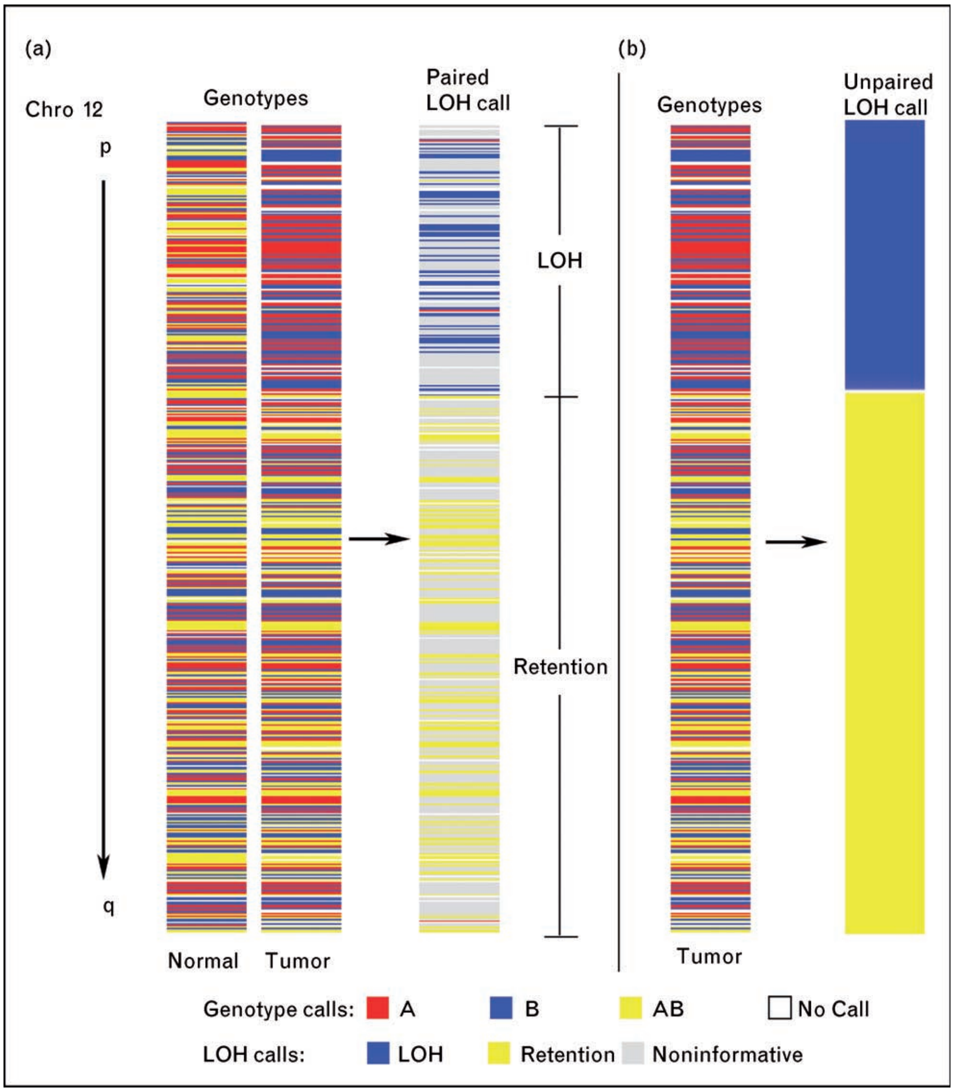
Advances in genotyping technology have enabled approximately five-fold yearly increases in the number of SNPs one can genotype in parallel in a single experiment, to the point that current methods support genotyping of over half a million SNPs across the genome simultaneously. The two major technologies in this regard involve oligonucleotide probes either spotted on gene chips (Affymetrix) or adsorbed on beads (Illumina). Affymetrix SNP arrays, adapted from the microarray hybridization chip technology utilized for gene expression studies, are commercially available in 10 000, 100 000 and 500 000 SNP loci format. These arrays are manufactured using a photolithography process and contain up to 40 separate 25-mer oligonucleotide probes for each SNP locus, representing both mismatch and perfect match probes. Genomic DNA is prepared by restriction digest followed by adapter ligation and single-primer polymerase chain reaction (PCR), to provide amplification of a reduced representation of the genome. After DNA labeling and hybridization, fluorescence intensities are measured for each allele of each SNP [10–12]. In contrast, on Illumina arrays allele-specific oligonucleotide probes are adsorbed to microbeads arranged on a microarray. DNA is prepared by Φ29-based whole-genome amplification, followed by fragmentation and hybridization to the array. Signal intensities are obtained by incorporating fluorescent nucleotides as these fragments are extended on the array. This technology is more amenable to custom-designed arrays [9••,13]. Predesigned arrays are also available to genotype 109 000, 240 000, 317 000 or 550 000 SNP loci.

These platforms generally require high-quality DNA from fresh or fresh frozen samples. Genotyping of DNA from formalin-fixed paraffin-embedded (FFPE) tissue has been performed in microarray format at lower SNP densities, including 1500 SNPs across the genome using Affymetrix arrays and DNA prepared by multiplex PCR [14–16], and 6000 SNPs across the genome using Illumina arrays without whole genome amplification [17]. The ability to robustly amplify lower-quality DNA from FFPE tissue, to amounts and levels of homogeneity that allow for a very high-resolution analysis, has, however, been elusive and continues to be the focus of ongoing research.

Determining LOH with and without paired normal DNA

LOH, the somatic conversion of heterozygous germline alleles to homozygosity, represents a key step in the inactivation of multiple TSGs [18]. Traditionally, loci harboring LOH have been determined by genotyping microsatellite markers. The higher density of SNP makers, however, enables detection of LOH at much higher resolution, and LOH mapping is increasingly being performed using high-resolution SNP arrays [14,19,20]. The most common and straightforward method is to directly compare genotypes between tumor and paired germline DNA (Fig. 2A). Only SNPs that are heterozygous in the germline are informative as to whether LOH has taken place. Various methods, such as Hidden Markov Model-based and counting methods, have been applied to define the boundaries of the regions undergoing LOH, based upon the distribution of SNPs in which LOH is directly observed [15]. To date, such analyses have been used to greatest effect to classify tumors and understand their progression. Classification schemes based upon LOH and hierarchical clustering have been used in the settings of lung cancer cell lines [21] and prostate cancer [14]. SNP arrays have been used to characterize LOH progression in samples from children with acute lymphoblastic leukemia who relapse after chemotherapy [22], suggest a common precursor to biphasic malignant components of Phylloides tumors

Figure 2 Determination of loss of heterozygosity from nucleotide polymorphism array data with and without paired normal data



Data from chromosome 12 from a single prostate tumor and its paired normal are presented. Each SNP locus is represented as a horizontal line, arranged in order from the p to q termini and colored according to the genotype or LOH call. (a) Direct comparison of data between the tumor and its paired normal reveals heterozygous loci in the normal that turn homozygous in the tumor (called LOH in the comparison view) and other heterozygous loci in the normal that remain heterozygous in the tumor (called 'retention' in the comparison view). Loci that are homozygous in the normal are noninformative in the comparison. The distribution of LOH and retention calls reveals that the p arm has undergone regional LOH whereas the q arm is retained. (b) The regions of LOH and retention in this chromosome can also be inferred from the unpaired tumor data, by consideration of the frequency of heterozygous calls in each region.

of the breast [20], and suggest that loss of imprinted loci in atypical adenomas may be an intermediate in the adenoma-carcinoma progression sequence in thyroid oncogenesis [23]. Moreover, SNP array-based analyses have shown that although LOH events are common in breast cancer tumor cells, they are infrequent in neighboring stromal elements [24].

An advantage of SNP arrays is that they provide marker densities that enable the identification of regions of LOH without the use of paired normal DNA [25•] (Fig. 2B). In this case, statistical analyses are applied to identify strings of consecutive homozygous SNPs that are longer than would be expected to appear by chance alone. As such, every SNP is informative; however, the resolution of such an analysis is necessarily lower than one can attain with paired normal data. Moreover, the haplotype block structure of the genome can lead to correlations between consecutive SNP genotypes, and therefore must be taken into account in the highest-density SNP array datasets [25•]. Such an approach is often necessary when paired germline DNA is not available (as in the case of most cell lines and xenografts). Moreover, because only tumors are being genotyped, genotyping costs are reduced by half.

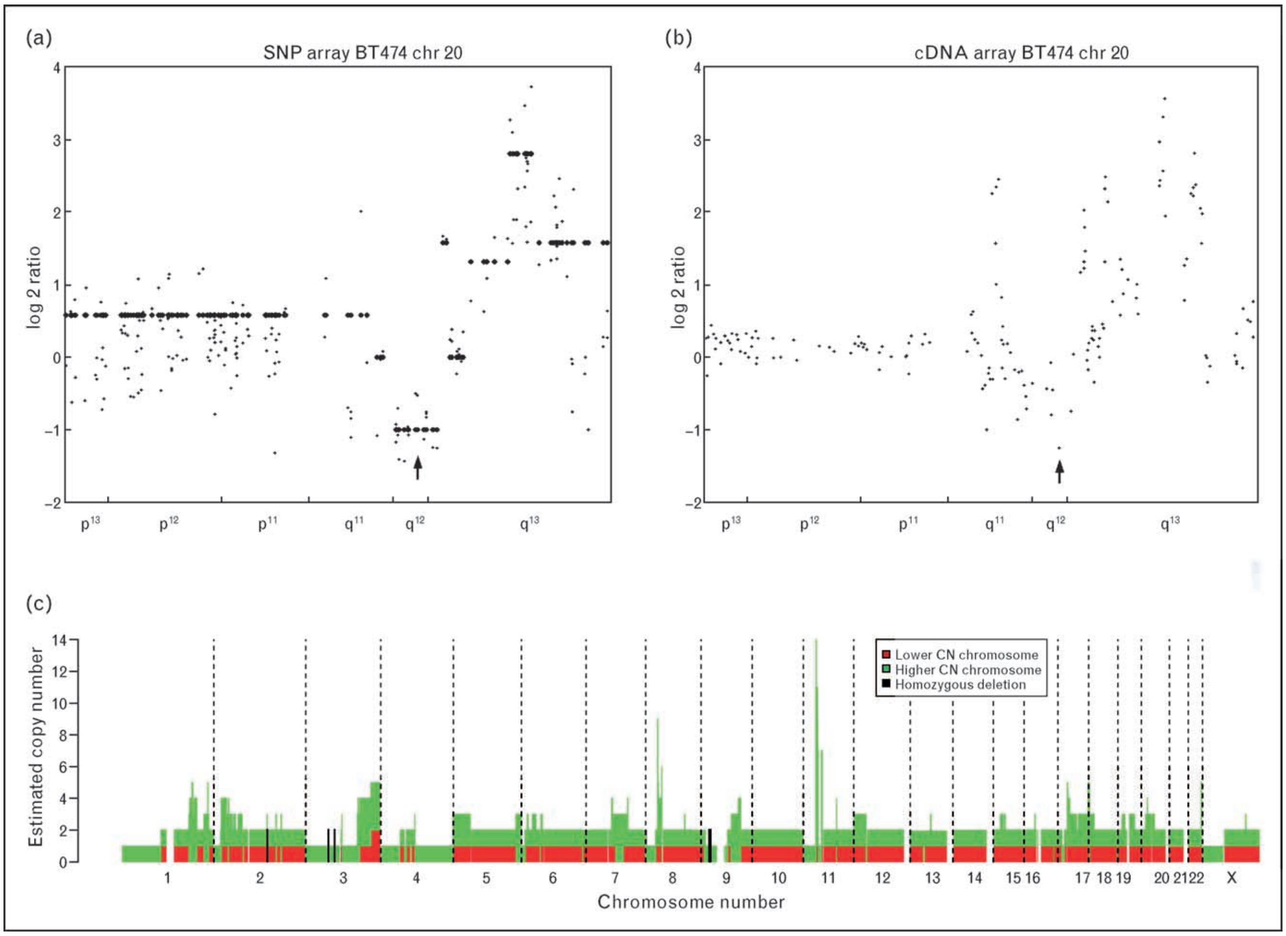
Similar approaches have also been used with pancreatic cancer cell lines to generate high-resolution allelotype and deletion breakpoint maps [23] and with acute myeloid leukemias to identify prevalent regions of LOH [26].

LOH is not always equivalent to copy number losses. LOH can arise due to hemizygous deletion alone, or followed by gene duplication leading to copy-neutral LOH. Conversely, loss of a single chromosome in a hyperploid cell may give the appearance of a deletion but leave two remaining chromosomes with retention of heterozygosity. Up to 80% of LOH events in some sample sets reflect copy-neutral LOH [25•]. LOH is also predominantly copy-neutral at particular loci in certain tumor types, such as acute myeloid leukemia [26], medulloblastoma [27], and basal cell carcinoma [28]. JAK2 is a specific target of copy-neutral LOH in myeloproliferative diseases [29]. Conversely, high-level amplifications of a single allele can lead to the inability to detect the second allele (allelic imbalance), giving the false appearance of LOH. Therefore, an integrative analysis of LOH and copy number changes is imperative.

Determining copy number variations on a genome-wide basis

DNA copy number variations can be determined from many SNP array datasets by comparing the hybridization signal intensities at each SNP locus with corresponding signal intensities from normal genomes [1,2] (Fig. 3). The largest body of work exists for Affymetrix arrays, where several algorithms now exist to combine probe-level signal intensities to an overall SNP-level signal intensity reflective of underlying copy number. The earliest algorithms to be developed were dChipSNP [30], which determines signal intensities using a model-based

Figure 3 Determination of overall and allele specific copy number variation



DNA copy number variations were determined by comparing locus-specific hybridization signal intensities in tumor genomes with corresponding signal intensities from normal genomes. (a) and (b) Log2 copy number ratios determined from 10K SNP arrays and cDNA-based comparative genomic hybridization arrays are displayed for chromosome 20 of the BT474 breast carcinoma cell line. Similar copy number changes were identified using both array types. The arrow denotes a hemizygous deletion that was further validated by quantitative PCR. Adapted from [2•]. (c) Parent-specific copy numbers across the entire genome are displayed for the HCC95 lung adenocarcinoma cell line. When the total copy number adds to more than 2, it is almost always seen to be due to an amplification of only one of the two alleles. Adapted from [41••].

method originally developed for analysis of expression arrays [31] and Copy Number Analysis Tool [32], which compares signal intensities at the perfect match probes to signal intensity ratios between perfect match and mismatch probes. More recently, algorithms have been developed to reduce noise levels introduced by variations in conditions between experiments. These include Copy Number Analyzer for GeneChip (CNAG) [33•], Genomic Imbalance Map (GIM) [34•], and Copy Number Analysis with Regression And Tree (CARAT) [35•]. All of these consider signal intensities at an SNP locus to be dependent not only on the underlying copy number of the sample, but also take into account how well restriction digested fragments of varying lengths and GC contents amplify, under varying experimental conditions, during DNA preparation for hybridization to the array.

With current SNP arrays interrogating over 500 000 loci across the genome, the resolution offered by these arrays matches or exceeds most state-of-the-art comparative genomic hybridization methods. Therefore, increasing numbers of tumors have been analyzed in this fashion. In prostate cancer, amplifications of TPD52 [36] and prosaposin [37] have been noted. In lung cancer, several regions of homozygous deletion and high-level amplification were observed [2]. In colon cancer, copy number changes were found to be associated with changes in gene expression [38]. The resolution afforded by these studies is beginning to allow identification of individual oncogenes and TSGs. For instance, MITF was shown to be an oncogene after amplifications were observed at this locus in an SNP array study of the NCI60 panel of cancer cell lines [39••]. More recently, amplifications of NOTCH3 were noted in ovarian tumors by an SNP array analysis, and the functional role of NOTCH3 was suggested by the ability to suppress cell proliferation by inhibiting NOTCH3 [40•].

Allele-specific copy number determination

In addition to determining overall copy number variations, the presence of signal intensity data corresponding to each of the two alleles allows the determination of allele-specific copy numbers. Several algorithms now exist to do this. CNAG, GIM and CARAT require paired normal data from the same individual to perform this function; probe-level allele-specific quantitation [41••] does not. The ability to determine allele-specific copy numbers at heterozygous SNP loci enables determination of parent-specific copy numbers over larger regions, namely, that whether an amplified or deleted chromosomal region is derived preferentially from one parent (Fig. 3C). In fact, we found that when a region is amplified, the extra copies tend to be derived from a single parent, rather than both alleles being equally amplified [41••]. This may reflect preferential amplification of an allele harboring an activating mutation. More intriguing is the possibility that certain germline variants are preferentially amplified across tumors, not least because these germline variants would be candidate cancer risk alleles for lending a predisposition to developing cancer.

Role of germline variants in cancer predisposition

A more straightforward method of identifying cancer risk alleles is to perform association studies to identify germline variants that segregate between populations with and without cancer. With SNP arrays reaching densities that enable whole-genome association studies, these arrays have now been used to identify common risk alleles for other diseases such as macular degeneration [42] and chronic lymphocytic leukemia [43]. These results have been especially interesting because they have implicated genes that appear to be involved in the development of these diseases [42]. The risk of developing cancer is known to have a genetic component, and common risk alleles for some cancers have already been identified using other genotyping methods [44]. With the high-throughput genotyping enabled by SNP arrays, we expect them to be used to identify many more cancer risk alleles in the near future. Identification of these alleles will not only mark who is at risk for specific types of cancer, but also point to genes that may be targeted for their roles in the development of cancer [45].

Conclusion

The current generation of SNP arrays interrogates ~25 SNPs per gene; this is expected to double in 2007 with the announcement by Affymetrix of the release of a 1-million SNP array. At these densities, we have reached a critical phase in SNP array analysis of cancer, where we have attained a resolution that should enable the identification of many more oncogenes and TSGs. Moreover, with the ability to simultaneously map regions of copy number changes and LOH in individual tumors, these arrays are likely to become invaluable tools for classifying individual cancers based upon a comprehensive characterization of the somatic genetic changes they have undergone.

The major challenges to the field range from data generation to analysis. Most importantly, the ability to generate robust data from FFPE will open up major opportunities to study large tumor archives with long-term clinical follow-up. The primary analytic challenge is presented by the sheer volume of data being generated, that increasingly requires highly automated tools to summarize these datasets and point to the most interesting findings. Furthermore, the opportunities offered by our newfound ability to determine allele-specific copy numbers have not yet been fully explored. A complete understanding of the cancer genome requires an integrated analysis of the multiple types of alterations that may accrue. With the ability to simultaneously characterize LOH, overall copy number changes, and allele-specific copy number changes, all at very high resolution, SNP arrays offer a unique platform for such an integrated analysis.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

Additional references related to this topic can also be found in the Current World Literature section in this issue (pp. 000–000).

1 Bignell GR, Huang J, Greshock J, *et al.* High-resolution analysis of DNA copy number using oligonucleotide microarrays. Genome Res 2004; 14:287–295.

2 Zhao X, Li C, Paez JG, *et al.* An integrated view of copy number and allelic alterations in the cancer genome using single nucleotide polymorphism arrays. Cancer Res 2004; 64:3060–3071.

3 Botstein D, Risch N. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. Nat Genet 2003; 33 (Suppl):228–237.

4 Kruglyak L, Nickerson DA. Variation is the spice of life. Nat Genet 2001; 27:234–236.

5 International HapMap Consortium. A haplotype map of the human genome. Nature 2005; 437:1299–1320.

6 Gabriel SB, Schaffner SF, Nguyen H, *et al.* The structure of haplotype blocks in the human genome. Science 2002; 296:2225–2229.

7 Nicolas P, Sun F, Li LM. A model-based approach to selection of tag SNPs. BMC Bioinformatics 2006; 7:303.

8 Cutler JA, Mitchell MJ, Greenslade K, *et al.* A rapid and cost-effective method for analysis of three common genetic risk factors for thrombosis. Blood Coagul Fibrinolysis 2001; 12:33–36.

9 •• Gunderson KL, Steemers FJ, Lee G, *et al.* A genome-wide scalable SNP genotyping assay using microarray technology. Nat Genet 2005; 37:549–554.
Describes the development of bead-based array for genome-wide SNP genotyping and copy number analysis.

10 Matsuzaki H, Dong S, Loi H, *et al.* Genotyping over 100,000 SNPs on a pair of oligonucleotide arrays. Nat Methods 2004; 1:109–111.

11 Hardenbol P, Baner J, Jain M, *et al.* Multiplexed genotyping with sequence-tagged molecular inversion probes. Nat Biotechnol 2003; 21:673–678.

12 Lockhart DJ, Dong H, Byrne MC, *et al.* Expression monitoring by hybridization to high-density oligonucleotide arrays. Nat Biotechnol 1996; 14:1675–1680.

13 Gunderson KL, Kruglyak S, Graige MS, *et al.* Decoding randomly ordered DNA arrays. Genome Res 2004; 14:870–877.

14 Lieberfarb ME, Lin M, Lechpammer M, *et al.* Genome-wide loss of heterozygosity analysis from laser capture microdissected prostate cancer using single nucleotide polymorphic allele (SNP) arrays and a novel bioinformatics platform dChipSNP. Cancer Res 2003; 63:4781–4785.

15 Lin M, Wei LJ, Sellers WR, *et al.* dChipSNP: significance curve and clustering of SNP-array-based loss-of-heterozygosity data. Bioinformatics 2004; 20:1233–1240.

16 Dumur CI, Dechsukhum C, Ware JL, *et al.* Genome-wide detection of LOH in prostate cancer using human SNP microarray technology. Genomics 2003; 81:260–269.

17 Lips EH, Dierssen JW, van Eijk R, *et al.* Reliable high-throughput genotyping and loss-of-heterozygosity detection in formalin-fixed, paraffin-embedded tumors using single nucleotide polymorphism arrays. Cancer Res 2005; 65:10188–10191.

18 Presneau N, Manderson EN, Tonin PN. The quest for a tumor suppressor gene phenotype. Curr Mol Med 2003; 3:605–629.

19 Lindblad-Toh K, Tanenbaum DM, Daly MJ, *et al.* Loss-of-heterozygosity analysis of small-cell lung carcinomas using single-nucleotide polymorphism arrays. Nat Biotechnol 2000; 18:1001–1005.

20 Wang ZC, Buraimoh A, Iglehart JD, Richardson AL. Genome-wide analysis for loss of heterozygosity in primary and recurrent phyllodes tumor and fibroadenoma of breast using single nucleotide polymorphism arrays. Breast Cancer Res Treat 2006; 97:301–309.

21 Janne PA, Li C, Zhao X, *et al.* High-resolution single-nucleotide polymorphism array and clustering analysis of loss of heterozygosity in human lung cancer cell lines. Oncogene 2004; 23:2716–2726.

22 Irving JA, Bloodworth L, Bown NP, *et al.* Loss of heterozygosity in childhood acute lymphoblastic leukemia detected by genome-wide microarray single nucleotide polymorphism analysis. Cancer Res 2005; 65:3053–3058.

23 Sarquis MS, Weber F, Shen L, *et al.* High frequency of loss of heterozygosity in imprinted, compared with nonimprinted, genomic regions in follicular thyroid carcinomas and atypical adenomas. J Clin Endocrinol Metab 2006; 91:262–269.

24 Allinen M, Beroukhim R, Cai L, *et al.* Molecular characterization of the tumor microenvironment in breast cancer. Cancer Cell 2004; 6:17–32.

25 • Beroukhim R, Lin M, Park Y, *et al.* Inferring loss-of-heterozygosity from unpaired tumors using high-density oligonucleotide SNP arrays. PLoS Comput Biol 2006; 2:e41.
Highly accurate method for determining LOH without paired normals.

26 Raghavan M, Lillington DM, Skoulakis S, *et al.* Genome-wide single nucleotide polymorphism analysis reveals frequent partial uniparental disomy due to somatic recombination in acute myeloid leukemias. Cancer Res 2005; 65:375–378.

27 Langdon JA, Lamont JM, Scott DK, *et al.* Combined genome-wide allelotyping and copy number analysis identify frequent genetic losses without copy number reduction in medulloblastoma. Genes Chromosomes Cancer 2006; 45:47–60.

28 Teh MT, Blaydon D, Chaplin T, *et al.* Genomewide single nucleotide polymorphism microarray mapping in basal cell carcinomas unveils uniparental disomy as a key somatic event. Cancer Res 2005; 65:8597–8603.

29 Kralovics R, Passamonti F, Buser AS, *et al.* A gain-of-function mutation of JAK2 in myeloproliferative disorders. N Engl J Med 2005; 352:1779–1790.

30 Paez JG, Lin M, Beroukhim R, *et al.* Genome coverage and sequence fidelity of phi29 polymerase-based multiple strand displacement whole genome amplification. Nucleic Acids Res 2004; 32:e71.

31 Li C, Wong WH. Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. Proc Natl Acad Sci USA 2001; 98:31–36.

32 Huang J, Wei W, Zhang J, *et al.* Whole genome DNA copy number changes identified by high density oligonucleotide arrays. Hum Genomics 2004; 1:287–299.

33 • Nannya Y, Sanada M, Nakazaki K, *et al.* A robust algorithm for copy number detection using high-density oligonucleotide single nucleotide polymorphism genotyping arrays. Cancer Res 2005; 65:6071–6079.
Developed improved overall copy number detection and allele-specific copy number detection methods.

34 • Ishikawa S, Komura D, Tsuji S, *et al.* Allelic dosage analysis with genotyping microarrays. Biochem Biophys Res Commun 2005; 333:1309–1314.
Developed improved overall copy number detection and allele-specific copy number detection methods.

35 • Huang J, Wei W, Chen J, *et al.* CARAT: a novel method for allelic detection of DNA copy number changes using high density oligonucleotide arrays. BMC Bioinformatics 2006; 7:83.
Developed improved overall copy number detection and allele-specific copy number detection methods.

36 Rubin MA, Varambally S, Beroukhim R, *et al.* Overexpression, amplification, and androgen regulation of TPD52 in prostate cancer. Cancer Res 2004; 64:3814–3822.

37 Koochekpour S, Zhuang YJ, Beroukhim R, *et al.* Amplification and overexpression of prosaposin in prostate cancer. Genes Chromosomes Cancer 2005; 44:351–364.

38 Tsafrir D, Bacolod M, Selvanayagam Z, *et al.* Relationship of gene expression and chromosomal abnormalities in colorectal cancer. Cancer Res 2006; 66:2129–2137.

39 •• Garraway LA, Widlund HR, Rubin MA, *et al.* Integrative genomic analyses identify MITF as a lineage survival oncogene amplified in malignant melanoma. Nature 2005; 436:117–122.
Identified and functionally validated the oncogene MITF.

40 • Park JT, Li M, Nakayama K, *et al.* Notch3 gene amplification in ovarian cancer. Cancer Res 2006; 66:6312–6318.
Identified amplification of NOTCH 3, which appears to play a functional role in ovarian cancer.

41 •• LaFramboise T, Weir BA, Zhao X, *et al.* Allele-specific amplification in cancer revealed by SNP array analysis. PLoS Comput Biol 2005; 1:e65.
This paper develops a method to determine allele-specific copy-numbers from SNP array data and shows amplifications tend to be of a single allele.

42 Klein RJ, Zeiss C, Chew EY, *et al.* Complement factor H polymorphism in age-related macular degeneration. Science 2005; 308:385–389.

43 Sellick GS, Webb EL, Allinson R, *et al.* A high-density SNP genomewide linkage scan for chronic lymphocytic leukemia-susceptibility loci. Am J Hum Genet 2005; 77:420–429.

44 Amundadottir LT, Sulem P, Gudmundsson J, *et al.* A common variant associated with prostate cancer in European and African populations. Nat Genet 2006; 38:652–658.

45 Sellick GS, Longman C, Tolmie J, *et al.* Genomewide linkage searches for Mendelian disease loci can be efficiently conducted using high-density SNP genotyping arrays. Nucleic Acids Res 2004; 32:e164.

CCO
Manuscript No. 115

Current Opinion in Oncology
Typeset by Thomson Digital
for Lippincott Williams & Wilkins

Dear Author,

During the preparation of your manuscript for typesetting, some queries have arisen. These are listed below. Please check your typeset proof carefully and mark any corrections in the margin as neatly as possible or compile them as a separate list. This form should then be returned with your marked proof/list of corrections to the Production Editor.

QUERIES: to be answered by AUTHOR/EDITOR

AUTHOR: The following queries have arisen during the editing of your manuscript. Please answer the queries by marking the requisite corrections at the appropriate positions in the text.

QUERY NO.	QUERY DETAILS	
	No Query.	